(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 September 2001 (13.09.2001)

PCT

(10) International Publication Number
WO 01/66804 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/US01/07775

(22) International Filing Date: 9 March 2001 (09.03.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/521,983 9 March 2000 (09.03.2000) US
09/613,517 10 July 2000 (10.07.2000) US

(71) Applicant: PROTOGENE LABORATORIES, INC. [US/US]; 303 Constitution Drive, Menlo Park, CA 94025 (US).

(72) Inventors: CRONIN, Maureen, T.; 771 Anderson Drive, Los Altos, CA 94024 (US). FRUEH, Felix; 511 Lakeview Way, Emerald Hills, CA 94062 (US). BRENNAN, Thomas, M.; 1998 Broadway #1505, San Francisco, CA 94109 (US).

(74) Agent: HALLUIN, Albert, P.; Howrey Simon Arnold & White LLP, 301 Ravenswood Avenue, Menlo Park, CA 94025 (US).

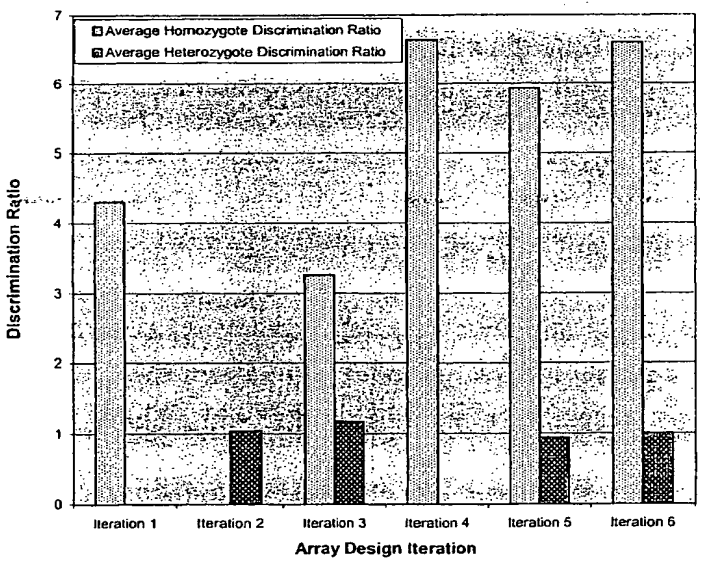
(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHODS FOR OPTIMIZING HYBRIDIZATION PERFORMANCE OF POLYNUCLEOTIDE PROBES AND LOCALIZING AND DETECTING SEQUENCE VARIATIONS



(57) Abstract: The present invention relates to a method for optimizing hybridization performance of polynucleotide probes on an array. More specifically, the present invention provides a cost-effective method for designing optimal polynucleotide probes and hybridization conditions to allow simultaneous determination of multiple sequence variations or multiple gene expression levels on an array under a single set of conditions. The present invention also relates to a method of localizing and detecting sequence variations. More specifically, the present invention provides a two-color system for sequence variation localization and detection. The present invention is applicable to high-throughput genotyping of known and unknown polymorphisms and mutations.

WO 01/66804 A2

METHODS FOR OPTIMIZING HYBRIDIZATION PERFORMANCE OF POLYNUCLEOTIDE PROBES AND LOCALIZING AND DETECTING SEQUENCE VARIATIONS

FIELD OF THE INVENTION

The present invention relates to a method for optimizing hybridization performance of polynucleotide probes on an array. More specifically, the present invention provides a cost-effective method for designing optimal polynucleotide probes and hybridization conditions to allow simultaneous determination of multiple sequence variations or multiple gene expression levels on an array under a single set of conditions. The present invention also relates to a method for localizing and detecting sequence variations. More specifically, the present invention provides a two-color system for sequence variation localization and detection. The present invention is applicable to high-throughput genotyping of known and unknown polymorphisms and mutations.

BACKGROUND OF THE INVENTION

Intense efforts are under way to map and sequence the human genome and the genomes of many other species. In February 2001, a draft sequence of the human genome was published (International Human Genome Sequencing Consortium, *Nature* 409:860-921 (2001) and Venter *et al.*, *Science*, 291:1304-1305 (2001)). This information, however, represents only a reference sequence of the 3-billion-base human genome. The remaining task lies in the determination of sequence variations (*e.g.*, mutations, polymorphisms, haplotypes) and sequence functions, which are important for the study, diagnosis, and treatment of human genetic diseases.

In addition to the human genome, the mouse genome is being sequenced. Genbank provides about 1.2% of the 3-billion-base mouse genome and a rough draft of the mouse genome is expected to be available by 2003 and a finished genome by 2005 (http://www.informatics.jax.org). The Drosophila Genome Project has also been completed recently (http://www.fruitfly.org). Thus far, genomes of more than 30 organisms have been sequenced (http://www.tigr.org and http://www.ncbi.nlm.nih.gov/genomes/index.html).

During the past decade, the development of array-based hybridization technology has received great attention. This high throughput method, in which

hundreds to thousands of polynucleotide probes immobilized on a solid surface are hybridized to target nucleic acids to gain sequence and function information, has brought economical incentives to many applications. See, *e.g.*, McKenzie, *et al., Eur. J. of Hum. Genet.* 6:417-429 (1998), Green *et al., Curr.Opin. in Chem. Biol.* 2:404-410 (1998), and Gerhold *et al., TIBS,* 24:168-173 (1999).

One application of the array technology is the genotyping of mutations and polymorphisms, also known as re-sequencing. Typically, sets of polynucleotide probes, that differ by having A, T, C, or G substituted at or near the central position, are fabricated and immobilized on a solid support. Fluorescently labeled target nucleic acids containing the expected sequences will hybridize best to perfectly matched polynucleotide probes, whereas sequence variations will alter the hybridization pattern, thereby allowing the determination of mutations and polymorphic sites. See, *e.g.*, Wang, D., *et al., Science* 280:1077-1082 (1998) and Lipshutz, R., *et al., Nature Genetics Supplement* 21:20-24 (1999), and U.S. Patent Nos. 5,858,659, 5,856,104, 5,871,928, and 5,968,740.

Another application is the monitoring of expression level to compare gene expression patterns. In one type of array, many gene-specific polynucleotide probes derived from the 3' end of RNA transcripts are spotted on a solid surface. This array is then probed with fluorescently labeled cDNA representations of RNA pools from test and reference cells. The relative amount of transcript present in the pool is determined by the fluorescent signals generated and the level of gene expression is compared between the test and the reference cell (also known as the two-color fluorescence analysis). See, *e.g.*, Duggan, D., *et al., Nature Genetics Supplement* 21:10-14 (1999), DeRisi, J., *et al., Science* 278:680-686 (1997), and U.S. Patent Nos. 5,800,992 and 6,040,138.

Another application of the array technology is the *de novo* sequencing of target nucleic acids by polynucleotide hybridization. For example, an array of all possible 8-mer polynucleotide probes may be hybridized with fluorescently labeled target nucleic acids, generating large amounts of overlapping hybridization data. The reassembling of this data by computer algorithm can determine the sequence of target nucleic acids. See, *e.g.*, Drmanac, S. *et al., Nature Biotechnology* 116:54-58 (1998), Drmanac, S. *et al. Genomics* 4:114-28 (1989), and U.S. Patent Nos. 5,202,231, 5,525,464, and 5,972,619.

A critical step in array-based hybridization technology is finding a condition where there is sufficient discrimination between perfect matches and mismatches. One problem is that for a particular target sequence, there is only one perfect match with a polynucleotide probe, while there are many possible end and internal mismatches. Unless the discrimination is very strong, there will be an inevitable background problem contributed by a large number of end and internal mismatches. Another problem is the sequence dependence of hybridization performance. G/C base pairs form three hydrogen bonds as opposed to two hydrogen bonds between A/T base pairs. Therefore, polynucleotide probes rich in G/C pairs will form more stable hybridization complex with target nucleic acids than A/T rich polynucleotides. If a more stringent condition is chosen that allows effective discrimination between perfect matches and mismatches in G/C rich sequences, many A/T rich sequences may not form enough hybridization complex to be detected, which leads to false negatives. Alternatively, if one chooses a less stringent condition to stabilize the weak A/T rich sequences, there will not be enough discrimination against mismatches in G/C rich sequences and many false positives will result.

In an array where hundreds to thousands of polynucleotide probes are immobilized, it is difficult to find a condition that will maximize the hybridization performance for all probes. Although in some cases, it is possible to measure hybridization under a plurality of conditions with varying stringency to enhance the hybridization performance for all probes, the polynucleotide probes tend to respond similarly to adjustments in assay stringency conditions. Thus varying hybridization conditions is limited in creating the necessary discrimination ratio for reliable detection. In addition, additional steps and varying conditions will undoubtedly add time and cost to the hybridization assay. There is a need in the art for an effective and cost-saving method for modulating and optimizing hybridization performance of polynucleotide probes on an array.

In addition, current probe design strategies for hybridization-based sequence variation detection on arrays focus on complex tiling methods, which lead to increased number of probes required for each sequence variation determination (*e.g.*, WO 98/41657). Therefore, the information generated per probe is reduced and the cost of detection increases. More importantly, in the applications where the identity of a sequence variation is already characterized (e.g. in clinical environment), it is not necessary to identify the specific sequence variation, but simply its presence or

absence. There is a need in the art for a fast and cost-effective method for detecting sequence variations.

SUMMARY OF THE INVENTION

The present invention provides an iterative method of optimizing hybridization performance of array-immobilized polynucleotide probes to analyze target nucleic acid sequences. This method is applicable to simultaneously determining multiple sequence variations or simultaneously monitoring multiple gene expressions in target nucleic acid(s) under a single set of conditions.

In general, the present method comprises the steps of: (a) obtaining an array wherein a set of polynucleotide probes designed specifically for each sequence variation or each gene is immobilized on the array; (b) hybridizing target nucleic acid(s) to the array-immobilized polynucleotide probes under a pre-determined condition; (c) determining the differences in hybridization between target nucleic acid(s) and the array-immobilized polynucleotide probes; (d) changing the melting temperature, length, sequence composition, or hybridization environment of at least one polynucleotide probe; and (e) repeating steps (a)-(d), if necessary, until the differences in hybridization between target nucleic acid(s) and array-immobilized polynucleotide probes simultaneously indicate the presence or absence of two or more sequence variations in target nucleic acid(s) or simultaneously indicate the expression levels of two or more genes under the pre-determined condition. In particular, the melting temperature of a polynucleotide probe may be estimated from a mathematical formula. The melting temperature may be changed by no more than about 15, 10 or 5 °C. The length of a polynucleotide probe may be changed by less than about 10, 5, or 2 nucleotides. Methods of changing the sequence composition of a polynucleotide probe may include changing the G/C content of the polynucleotide probe, incorporating of a polynucleotide analog, among others. Methods of changing the hybridization environment may include using a chemical reagent such as a hybridization optimization reagent, a denaturing reagent, a chaotropic salt, and a renaturation accelerant, changing the linker molecule, changing the surface conditions, changing local concentrations of target nucleic acid(s) or polynucleotide probes, or applying electric current, among others. The sequence variations may be polymorphic forms or mutations, such as polymorphic forms or mutations of a gene, a

regulatory sequence, or an intronic sequence. The gene expression profiled may be a pool of RNAs or complementary DNAs or RNAs.

Further, the present invention provides an array wherein the melting temperatures of polynucleotide probes immobilized on the array differ by no more than about 15, 10, or 5 °C. The melting temperatures of polynucleotide probes may be estimated from a mathematical formula. The length of array-immobilized polynucleotide probes may differ by about 10, 5, or 2 nucleotides. The melting temperatures of polynucleotide probes immobilized on the array may also be within 10, 9, 8, 7, 6, or 5 °C of the average melting temperature.

Simultaneous determination of at least about 2, 5, 10, 50, 100, 1000, or 10,000 sequence variations may be performed on a single array. Typically, the density of polynucleotide probes on the array is between about 2-10,000 per cm^2, preferably lower than about 5,000, 2,000, 1,000, 400, or 100 per cm^2. Each polynucleotide probe may be about 6 to 100 nucleotides long, *e.g.*, shorter than about 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, or 90 nucleotides long. In the case of overlapping polynucleotides probes, the overlap may be about 1 to 50 bases, preferably below 30, 20, 10, or 5 bases.

The present invention also features a method for determining the presence or absence of a sequence variation in a target nucleic acid sequence comprising the steps of: (a) immobilizing at least two polynucleotide probes on a solid support wherein at least one polynucletide probe spans the location of the sequence variation; (b) attaching the target nucleic acid sequence with a first detectable label; (c) attaching a control nucleic acid sequence with a second detectable label wherein the second detectable label is different than the first detectable label; (d) contacting the immobilized polynucleotide probes with the mixture of the control nucleic acid sequence and the target nucleic acid sequence under hybridization conditions; and (e) determining the presence or absence of the sequence variation in the target nucleic acid sequence based on the hybridization pattern differences of polynucleotide probes.

The immobilization of polynucleotide probes on an array may be covalent or non-covalent. The polynucleotide probes may be synthesized *in situ* or presynthesized prior to the immobilization on the surface of an array. The *in situ* synthesis of polynucleotide probes may be performed on functionalized sites of an array. For example, array surface may be fabricated such that solutions on functionalized sites may be separated by surface tension. The area of each

functionalized site may be about 0.1 x 10^{-5} to 0.1 cm^2, preferably less than about 0.05, 0.01, or 0.005 cm^2. Typically, the total number of functionalized sites on an array is between about 10-500,000, preferably, less than about 100,000, 50,000, 10,000, 5000, 1000, 500, or 100. The *in situ* synthesis may be performed using an ink jet printer apparatus, such as a piezoelectric pump.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates the global homozygote and heterozygote discrimination ratio values for each NAT2 genotyping array design.

Figures 2A-2B illustrate a detailed example of probe set optimization for T341C polymorphism. Figure 2A shows a typical hybridization to the constant length probe set in the first array design. Figure 2B shows a typical hybridization to the third array design which has Tm matched probes averaging 64°C.

Figure 3 compares hybridization results using the fully optimized array for two patient samples, one that is heterozygous for the T341C polymorphism and one that is homozygous for T at that site.

Figure 4 shows signals obtained for β-actin probes chosen at starting positions 335 (left, probe 1) and 600 (right, probe 2). Probes are 45 base pairs in length. Probe 1 (left) produces a significantly less intense signal than probe 2 (right).

Figure 5 shows probes for β-actin selected at different starting locations (indicated as number below bars). Probes 1 and 2 are represented as black bars. Bars represent intensities and are expressed as percentage of the most intense signal (obtained for probe 1025+).

Figure 6 shows the influence of probe length for probes selected at three different starting positions. Probe length is indicated in base pairs. Intensities for 45mer probes are in agreement with numerical data shown in Example 10. PM: perfect matched probes. 3 MM, 5 MM: three, five mismatches introduced in the center of the probes, respectively.

Figure 7 illustrates an example of designing overlapping polynucleotide probes for detecting sequence variations.

Figures 8 illustrates the mixing of differently labeled control nucleic acids with target nucleic acids for hybridization with array-immobilized polynucleotide probes. The star sign indicates the location of a sequence variation.

Figure 9 shows the hybridization results of a sequence variation detection using two-color fluorescent analysis.

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a method for optimizing hybridization performance of polynucleotide probes on an array. More specifically, the present invention provides a method for designing optimal polynucleotide probes and hybridization conditions to allow simultaneous determination of two or more sequence variations or two or more gene expression levels on a single array under a single set of conditions. The present invention also provides a method for cost effective iteration of array designs necessary to evaluate hybridization performance of a large number of polynucleotide probes. The present invention is applicable to genotyping of known polymorphisms and mutations, profiling gene expression levels, and identifying previously unknown nucleotide sequences. The present invention maximizes the information yield of hybridization-based array applications by increasing the number of informative array-immobilized polynucleotide probes.

The present invention also relates to a method for localizing and detecting sequence variations. More specifically, the present invention provides a two-color system for sequence variation localization and detection. The present invention is applicable to high-throughput genotyping of known and unknown polymorphisms and mutations.

In array based technology, target nucleic acids are determined by analyzing the extent of hybridization between the target sequence and polynucleotide probes on an array. The fundamental aspect of this technology is the discrimination of hybridization stability between the match and the mismatch. A problem to this discrimination is that a perfect match in A/T rich hybridization complexes would often have a lower stability than a mismatch in G/C rich hybridization complexes. This dependency of stability on base composition may lead to false positives if the stringency of hybridization conditions is low (*e.g.*, low hybridization temperature), as mismatches in G/C rich hybridization complexes may be stabilized and may behave like perfect matches. In contrast, when the stringency of hybridization conditions is high (*e.g.*, high hybridization temperature), false negatives may occur, as perfect matches between the A/T rich sequences may not form stable hybridization complexes. Therefore, for successful and reliable determination of target nucleic

acids, optimization of hybridization performance of polynucleotide probes is an essential step.

Elaborate methods of probe design and *in situ* polynucleotide synthesis have been developed. See, *e.g.*, U.S. Patent Nos. 5,695,940, 5,856,104, and 5,858,659; PCT publications WO 98/31836A1, WO 97/10365A1, WO 95/11995, and WO 97/2317, all incorporated herein by reference. These methods however present many problems to the modulation of hybridization performance of polynucleotide probes on an array. First, probe designs employing the tiling strategy or using all possible short polynucleotides create a large number of polynucleotide probes with a wide range of hybridization stability. For example, in the latter case, polynucleotide probes as many as 65,536 possible 8-mers or 262,144 possible 9-mers are examined for hybridization performance. It is difficult to modulate the stringency of hybridization conditions such that hundreds to thousands of probes will exhibit similar hybridization behaviors. Second, due to the low chemical coupling yield of *in situ* polynucleotide synthesis using photolithography, each probe site may contain a substantial number of truncated polynucleotide probes in addition to the desired full length probes. For example, in 10-mer and 20-mer probe sites, only about 40% and 15% of the polynucleotide probes are of the full length respectively (Forman, J., *et al.*, Molecular Modeling of Nucleic Acids, Chapter 13, pp 206-228, American Chemical Society (1998)) and McGall *et al.*, *J. Am. Chem. Soc.*, 119:5081-5090 (1997)). This probe length heterogeneity inevitably leads to unpredictable hybridization performance of polynucleotide probes on an array. Third, it is often necessary to introduce polynucleotide analogs to balance the stability difference between A/T rich and G/C rich sequences. Incorporation of unnatural structures into the polynucleotide probes in photolithography method involves new photodeprotection chemistry and will likely encounter low yields. Fourth, iteratively changing the polynucleotide probe length as a function of its base composition in photolithography is technically complex and impractical. Cost for redesigning probes with different length and sequence is prohibitively high. Finally, current probe optimization strategies focus on more complex tiling methods, which may lead to increased number of probes required for each sequence variation determination (*e.g.*, WO 98/41657). Therefore, the information generated per probe is reduced on average and the cost of detection increases.

In order for array-based hybridization technology to gain widespread acceptance in commercial areas, it is necessary to develop a method for designing probes by modulating hybridization performance of polynucleotide probes and a method for fabricating new designs of polynucleotide probes in a rapid and cost effective manner. In a system where more than one sequence variation or more than one gene expression levels, it becomes even more important to modulate the hybridization behaviors of polynucleotide probes. It is desirable, for reasons of simplicity and economy, that the array-based hybridization be performed under a single set of conditions to detect multiple sequence variations or profile multiple gene expressions. This requires the coordination of hybridization performance of large numbers of polynucleotide probes under a specific set of conditions to simultaneously probe two or more sequence variations or two or more gene expression levels in target nucleic acids.

In general, the present invention involves designing a first set of polynucleotide probes, which are immobilized on an array. This initial probe set may include probes complementary to the reference sequences. The initial probe set may also include control probes. The reference sequences are specific for each sequence variation to be determined or each gene to be profiled. Multiple sequence variations to be determined in a target nucleic acid may represent known variants of the reference sequence at different locations. A target nucleic acid may then be hybridized to the array-immobilized probes. The relative hybridization intensities of the probes to the target nucleic acid are determined and analyzed to estimate the presence or absence of each sequence variant or the level of each gene expression in the target nucleic acid. In order to simultaneous determine multiple sequence variations or multiple gene expressions, a second probe set may be designed wherein the hybridization performance of one or more polynucleotide probes are modified. In particular, melting temperature analysis may be performed. Probe length, composition or hybridization environment may be altered to improve the hybridization performance of polynucleotide probes for simultaneous detection. In particular, the second probe set may have less differentiation in melting temperatures. For example, the melting temperatures of polynucleotide probes may differ by less than about 10, 5, or 2°C. The melting temperatures of polynucleotide probes immobilized on the array may also be within 10, 9, 8, 7, 6, or 5 °C of the average melting temperature. The second set of polynucleotide probes may then be

immobilized on an array. The target nucleic acid is hybridized to the second set of array-immobilized probes and the relative hybridization of the probes to the target nucleic acid is determined. Each sequence variant or gene expression of the target nucleic acid is then reestimated from the relative hybridization intensities of the probes. The cycles of melting temperature evaluation and probe set design can be reiterated, if desired, until all sequence variations or gene expression levels of the target nucleic acid may be determined simultaneously under a singe set of conditions.

I. Initial design of polynucleotide probes

In one aspect, the present invention is suitable for determining precharacterized polynucleotide sequence variations. In other words, the genotyping is performed after the location and nature of polymorphic forms or mutations have already been determined. The sequences of known polymorphic forms, the wild-type/mutation sequences, and gene sequences may be referred to as reference sequences. For example, the two polymorphic forms of a biallelic single nucleotide polymorphism (SNP) may be used as two reference sequences. To analyze a deletion mutation, one can select the wild-type form and the deleted form as two reference sequences. In some instances, sequence variations of both the coding and noncoding strands of the target nucleic acid sequence may be determined. Therefore, both the coding and noncoding strands may be used as reference sequences for sequence variation determinations.

A substantial number of mutations and polymorphic forms have been reported in the published literature or may be accessible through publicly available web sites, such as from the draft human genome sequence (International Human Genome Sequencing Consortium; *Nature* 409:860-921 (2001); Genbank (http://www.ncbi.nlm.nih.gov), http://shgc.stanford.edu; http://www.tigr.org, among others. See also, Gelfand *et al.*, *Nucleic Acids Res.* 27:301-302 (1999) and Buetow *et al.*, *Nat. Genet.* 21:323-325 (1999). The availability of reference sequence information allows an initial set of polynucleotide probes to be designed for the identification of the known sequence variations.

The determination of sequence variations using the present invention also includes *de novo* characterizing polynucleotide sequence variations. In other words, genotyping may be used to identify points of new variations and the nature of new variations. For example, by analyzing a group of individuals representing ethnic

diversity among humans, the consensus or alternative alleles/haplotypes of the locus may be identified, and the frequencies in the population may be determined. Allelic variations and frequencies may also be determined for populations characterized by criteria such as geography, race, gender, among others. Such analysis may also be performed among different species in plants, animals, and other organisms. Examples of determining sequence variations can be found in U.S. Patent Nos. 5,858,659, 5,871,928 and PCT applications WO 98/56954, 98/38846, 99/14228, 98/30883, all incorporated herein by reference.

The present invention involves designing an initial set of polynucleotide probes based on reference sequences for each sequence variation. The reference sequences serve as a first estimate of the sequence variations in the target nucleic acid. The initial design of a probe set typically includes probes that are perfectly complementary to the reference sequences and span the location of each sequence variation. Perfect complementary means sequence-specific base pairing which includes *e.g.*, Waston-Crick base pairing or other forms of base pairing such as Hoogstein base pairing. In some instances, a series of overlapping polynucleotide probes perfectly complementary to the reference sequence may be employed. Leading or trailing sequences flanking the segment of complementarity can also be present. For example, a pair of polynucleotide probes perfectly complementary to the two polymorphic forms of a biallelic SNP (two reference sequences) may be employed. Of course, additional related polynucleotide probes may be added to improve the accuracy of the detection. More complex design of polynucleotide probes known to those skilled in the art may also be employed. For example, various tiling methods (*e.g.*, sequence tiling, block tiling, 4 x 3 tiling, and opt-tiling) are described in WO 95/11995, WO 98/30883, WO 98/56954, EP 717113A2, and WO/99/39004, all incorporated herein by reference.

A mismatch is when a sequence is not perfectly complementary to a reference sequence. Under suitable hybridization conditions, the perfectly matched would be expected to hybridize with its target sequence, but mismatch probes would not hybridize or would hybridize to a significantly lesser extent. Although one or more mismatches may be located anywhere in the mismatch probe, probes are often designed to have the mismatch locate at or near the center of the probe such that the mismatch is most likely to destabilize the hybridization complex with the target sequence. In addition, the mismatch site is typically not the location of the sequence

variation to be determined, but is within several nucleotides (*e.g.*, less than 5) on the 5' or 3' side of the sequence variation location. For example, a probe set for a known biallelic SNP may contain two groups of mismatch probes based on two reference sequences constituting the respective polymorphic forms. Each group of mismatch probes may include at least two sets of probes, which each set contains a series of probes with a mismatch at one nucleotide 5' and 3' to the polymorphic site.

The polynucleotide probe set may also include control probes. One class of control probes is normalization probes which provide a control for variation in hybridization condition, signal intensity, and other factors that may cause the signal of a perfect hybridization to vary between arrays. Typically, normalization probes are perfectly complementary to a known polynucleotide sequence that is added to the target nucleic acids. Normalization probes may be located throughout the array to control for spatial variation in hybridization intensity.

In a second aspect, the instant invention may be used to monitor and profile multiple gene expressions. The simultaneous monitoring of the expression levels of a multiplicity of genes permits comparison of relative expression levels and identification of biological conditions (*e.g.*, disease detection, drug screening, toxicology profiling) characterized by alterations of relative expression levels of various genes. The simultaneous monitoring of the expression levels also includes the determination of the presence or absence of genes.

Polynucleotide probes for expression monitoring may include probes each having a sequence that is complementary to a subsequence of one of the genes (or the mRNA or the corresponding antisense cRNA). The gene intron/exon structure and the relatedness of each probe to other expressed sequences may also be considered. Polynucleotide probe set may additionally include mismatch controls, normalization probes, among others. In particular, normalization probes may include probes hybridize specifically with constituively expressed genes in the biological sample, such as β-actin, the transferrin receptor gene, the GAPDH gene, and the like. Examples of monitoring gene expression levels are shown in U.S. Patent Nos. 5,811,231, 5,965,352, 6,040,138, and 6,146,830, WO 01/06013, WO 01/05935, WO 00/71161, WO 00/58521, WO 00/58520, Lockhart *et al.*, *Nature* 405:827-836 (2000), Roberts *et al.*, *Science* 287:873-880 (2000), Hughes *et al.*, *Nature Genetics* 25:333-337 (2000), Hughes *et al.*, *Cell* 102:109-126 (2000), Duggan, D., *et al.*, *Nature*

Genetics Supplement 21:10-14 (1999), and DeRisi, J., *et al., Science* 278:680-686 (1997), all incorporated herein by reference.

The number of polynucleotide probes for a sequence variation or a gene expression may vary depending on the nature of sequence variation, gene expression, and level of resolution desired. At least about 2, 5, 10, 20, 50, or 100 polynucleotide probes may be employed for each sequence variation or each gene. Simultaneous determination of at least about 2, 5, 10, 50, 100, 1000, or 10,000 sequence variations may be performed on a single array. Simultaneous profiling of at least about 2, 5, 10, 50, 100, 1000, or 100, 000 gene expressions may be performed on a single array. Each probe in both sequence variation determination and gene profiling may be about 6 to 100 nucleotides long, *e.g.* shorter than about 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, or 90 nucleotides long.

II. Array fabrication and immobilization of polynucleotide probes

Any suitable solid supports may be used in the present invention. These materials include glass, silicon, wafer, polystyrene, polyethylene, polypropylene, polytetrafluorethylene, among others. One of skill in the art will appreciate that there are many ways of immobilizing polynucleotides directly on an array (covalently or noncovalently), anchoring them to a linker moiety, or tethering them to an immobilized moiety. These methods are well taught in the art of solid phase synthesis (Protocols for oligonucleotides and analogs; synthesis and properties, *Methods Mol. Biol.* Vol. 20 (1993), incorporated herein by reference). The immobilization methods generally fall into one of the two categories: spotting of presynthesized polynucleotides and *in situ* synthesis of polynucleotides.

In the first category, preprepared polynucleotides are deposited onto known finite areas on an array. For example, traditional solid phase polynucleotide synthesis on controlled-pore glass (CPG) may also be employed and then simply printing presynthesized polynucleotides onto the array using direct touch or fine micropipetting. Polynucleotides may be synthesized on an automated DNA synthesizer, for example, on an Applied Biosystems synthesizer using 5-dimethoxytritylnucleoside β-cyanoethyl phosphoramidites. Synthesis of relatively long polynucleotide sequences may be achieved by PCR-based and/or enzymatic methods for economical advantages. Polynucleotides may be purified by gel

electrophoresis, HPLC, or other suitable methods known in the art before they are spotted or deposited on the solid support. Typical non-covalent linkages may include electrostatic interactions, ligand-protein interactions (*e.g.*, biotin/streptavidin or avidin interaction), and base-specific hydrogen bonding (*e.g.*, complementary base pairs), among others. For example, solid supports may be overlaid with a positively charged coating, such as amino silane or polylysine and presynthesized probes are then printed directly onto the solid surface. Printing may be accomplished by direct surface contact between the printing reagents and a delivery mechanism. The delivery mechanism may contain the use of tweezers, pins or capillaries, among others that serve to transfer polynucleotides or reagents to the surface. A variation of this simple printing approach is the use of controlled electric fields to immobilize prefabricated charged polynucleotides to microelectrodes on the array (*e.g.*, U.S. Patent No. 5,929,208 and WO 99/06593). For example, biotinylated polynucleotide probes may be directed to individual spots by polarizing the charge at that spot and then anchored in place via a steptavidin-containing permeation layer that covers the surface (Sosnowski *et al.*, *Proc. Natl. Acad. Sci.* 94:1119-1123 (1997) and Edman *et al.*, *Nucleic Acid. Res.* 25:4907-4914 (1997)). Some of the advantages of spotting technologies include ease of prototyping and therefore rapid implementation, low cost and versatility. In addition, presynthesized polynucleotides may be covalently attached to the solid surface, for example, using the method described in U.S. Patent No. 5,858,653.

In the second category, polynucleotides may be prepared by *in situ* synthesis on the array in a step-wise fashion. With each round of synthesis, nucleotide building blocks may be added to growing chains until the desired sequence and length are achieved in each spot. In general, *in situ* polynucleotide synthesis on an array may be achieved by two general approaches. First, photolithography may be used to fabricate polynucleotide on the array. For example, a mercury lamp may be shone through a photolithograhic mask onto the array surface, which removes a photoactive group, resulting in a 5' hydroxy group capable of reacting with another nucleoside. The mask therefore predetermines which nucleotides are activated. Successive rounds of deprotection and chemistry result in polynucleotides with increasing length. This method is disclosed in, *e.g.*, U.S. Patent Nos. 5,143,854, 5,489,678, 5,412,087, 5,744,305, 5,889,165, and 5,571,639, all incorporated herein by reference.

The second approach is the "drop-on-demand" method, which uses technology analogous to that employed in ink-jet printers (U.S. Patent Nos. 5,474,796, 5,985,551, 5,927,547, 6,177,558, Blanchard *et al.*, *Biosensors and Bioelectronics* 11:687-690 (1996), Schena *et al.*, *TIBTECH* 16:301-306 (1998), Green *et al.*, *Curr. Opin. in Chem. Biol.* 2:404-410 (1998), and Singh-Gasson, *et al.*, *Nat. Biotech.* 17:974-978 (1999), all incorporated herein by reference). This approach typically utilizes piezoelectric or other forms of propulsion to transfer reagents from miniature nozzles to solid surfaces. For example, the printer head travels across the array, and at each spot, electric field contracts, forcing a microdroplet of reagents onto the array surface. Following washing and deprotection, the next cycle of polynucleotide synthesis is carried out. The step yields in piezoelectric printing method typically equal to, and even exceed, traditional CPG polynucleotide synthesis. The drop-on-demand technology allows high-density gridding of virtually any reagents of interest. It is also easier using this method to take advantage of the extensive chemistries already developed for polynucleotide synthesis, for example, flexibility in sequence designs, synthesis of polynucleotide analogs, synthesis in the 5'-3' direction, among others. Because ink jet technology does not require direct surface contact, piezoelectric delivery is amendable to very high throughput production. Similar methods of reagent delivery using a tip of a spring probe are described in WO 99/05308, incorporated herein by reference.

In preferred embodiments, a piezoelectric pump may be used to add reagents to the *in situ* synthesis of polynucleotides. Microdroplets of 50 picoliters to 2 microliters of reagents may be delivered to the array surface. The design, construction, and mechanism of a piezoelectric pump are described in U.S. Patent Nos. 5,474,796 and 5,985,551. The piezoelectric pump may deliver minute droplets of liquid to a surface in a very precise manner. For example, a picopump is capable of producing picoliters of reagents at up to 10,000 Hz and accurately hits a 250 micron target at a distance of 2 cm.

Surface tension arrays (see, *e.g.*, U.S. Patent Nos. 5,474,796 and 5,985,551) may be employed in the present invention. Surface tension arrays are typically comprised of patterned hydrophilic and hydrophobic sites. A surface tension array may contain large numbers of hydrophilic sites against a hydrophobic matrix or vice versa, large numbers of hydrophobic sites against a hydrophilic matrix. A hydrophilic site typically includes free amino, hydroxyl group, as well as modified forms thereof,

such as activated or protected forms. A hydrophobic site typically includes alkyl, alkoxy, halide group. A hydrophobic site is typically inert to conditions of *in situ* synthesis. In surface tension arrays, a hydrophilic site is spatially segregated from neighboring hydrophilic sites because of the hydrophobic sites between hydrophilic sites. This spatially addressable pattern enables the precise and reliable location of chemicals or biologicals. The free amino, hydroxyl group of the hydrophilic sites may then be covalently coupled with a linker moiety capable of supporting chemical and biological synthesis. The hydrophilic sites may also support non-covalent attachment to chemicals or biologicals. Reagents delivered to the array are constrained by surface tension difference between hydrophilic and hydrophobic sites.

There are significant advantages to using surface tension arrays. The lithography and chemistry used to pattern the substrate surface are generic processes that simply define the array feature size and distribution. They are completely independent from the polynucleotide sequences that are synthesized or delivered at each site. In addition, the polynucleotide synthesis chemistry uses standard rather than custom synthesis reagents. The combined result is complete design flexibility both with respect to the sequences and lengths of polynucleotides used in the array, the number and arrangement of array features, and the chemistry used to make them. This method provides an inexpensive, flexible, and reproducible method for array fabrication.

Typically, the density of polynucleotide probes on the array is between about 2-10,000 per cm^2, preferably lower than about 5,000, 2,000, 1,000, 400, or 100 per cm^2. Each polynucletide probe may be about 6 to 100 nucleotides long, *e.g.* shorter than about 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, or 90 nucleotides long. In the case of overlapping polynucleotides probes, the overlap may be about 1 to 50 bases, preferably below 30, 20, 10, or 5 bases.

Typically, polynucleotide probes may be covalently or noncovalently attached to functionalized sites on a solid support. Functionalized sites are modifications of a solid support surface (e.g., hydrophilic sites, *infra*) for anchoring the *in situ* synthesis of polynucleotides or for supporting covalent or noncovalent attachment of the presynthesized polynucleotides. The area of each functionalized site may be about 0.1×10^{-5} to 0.1 cm^2, preferably less than about 0.05, 0.01, or 0.005 cm^2. Typically, the total number of functionalized sites on an array is between about 10-500,000, preferably, less than about 100,000, 50,000, 10,000, 5000, 1000, 500, or 100.

III. Preparation of target nucleic acids

The target nucleic acids may be prepared from human, animal, viral, bacterial, fungal, or plant sources using known methods in the art. For example, target sample may be obtained from an individual being analyzed. For assay of genomic DNA, virtually any biological sample is suitable. For example, convenient tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The target nucleic acids may also be obtained from other appropriate source, such as cDNAs, chromosomal DNA, microdissected chromosome bands, cosmid or YAC inserts, and RNA. Target nucleic acids may also be prepared as clones in M13, plasmid or lambda vectors and/or prepared directly from genomic DNA or cDNA. Examples of target nucleic acid preparation are described in *e.g.*, WO 97/10365.

The target nucleic acids are usually amplified, *e.g.*, by PCR prior to or during the detection of sequence variations. See, *e.g.*, *PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992). Primers may be selected to flank the borders of the sequence of interest. Suitable amplification methods also include the ligase chain reaction (LCF) (see Wu and Wallace, *Genomics* 4, 560 (1989), Landegren *et al.*, *Science* 241, 1077 (1988), transcription amplification (Kwoh *et al.*, *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication (Guatelli *et al.*, *Proc. Nat. Acad. Sci. USA,* 87, 1874 (1990)) and nucleic acid based sequence amplification (NASBA).

The target may be preferably fragmented before application to the array to reduce or eliminate the formation of secondary structures in the target. The fragmentation may be performed using a number of methods, including enzymatic, chemical; thermal cleavage or degradation. For example, fragmentation may be accomplished by heat/Mg^{2+} treatment, endonuclease (*e.g.*, DNAase 1) treatment, restriction enzyme digestion, shearing (*e.g.*, by ultrasound) or NaOH treatment.

It will be appreciated by one of skill in the art that the target nucleotide acids or the immobilized polynucleotide probes may be tagged with detectable labels. The labeling may occur before, during, or after hybridization, although in preferred embodiments, the target nucleic acids are labeled before hybridization. Detectable labels include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels may include biotin for staining with labeled streptavidin conjugate, magnetic beads

(*e.g.*, Dynabeads™), fluorescent molecules (*e.g.*, fluorescein, texas red, rhodamine, green fluorescent protein, FAM, JOE, TAMRA, ROX, HEX, TET, Cy3, C3.5, Cy5, Cy5.5, IRD41, BODIPY and the like), radiolabels (*e.g.*, ^{3}H, ^{251}I, ^{35}S, ^{34}S, ^{14}C, ^{32}P, or ^{33}P), enzymes (*e.g.*, horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA), colorimetric labels such as colloidal gold or colored glass or plastic (*e.g.*, polystyrene, polypropylene, latex, etc.) beads, mono and polyfunctional intercalator compounds.

Means of detecting such labels are also well known to those of skill in the art. For example, radiolabels may be detected using photographic film or scintillation counters. Fluorescent markers may be detected using a photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label.

IV. Hybridization between array-immobilized polynucleotide probes and target nucleic acids

Hybridization assays typically involve a hybridization mixture containing the target nucleic acids and other suitable reagents being brought into contact with the polynucleotide probes on the array and incubated at a temperature and for a time appropriate to allow hybridization between the target and polynucleotide probes. Usually, unhybridized target molecules may then be removed from the array by washing with a wash mixture that does not contain the target nucleic acids, such as a hybridization buffer. This leaves only hybridized target molecules. A predetermined condition for simultaneous determination of multiple sequence variations or gene expressions may be specified by temperature, concentration of reagents, hybridization and washing times, buffer components, and their pH and ionic strength, among others.

The hybridization can take place in any suitable container. Generally, incubation may be at temperatures normally used for hybridization of nucleic acids, for example, between about 20 °C and about 75 °C, *e. g.*, above about 30 °C, 40 °C, 50 °C, 60 °C, or 70 °C. The target nucleic acid may be incubated with the array for a time sufficient to allow the desired level of hybridization between the target and any complementary probes in the array, usually in about 10 minutes to several hours. But it may be desirable to hybridize longer, *e.g.*, overnight. After incubation with the

hybridization mixture, the array is usually washed with the hybridization buffer. Then the array may be examined to identify the polynucleotide probes to which the target has hybridized.

Suitable hybridization conditions may be determined by optimization procedures or experimental studies. Such procedures and studies are routinely conducted by those skilled in the art. See *e.g.*, Ausubel *et al.*, *Current Protocols in Molecular Biology, Vol. 1-2,* John Wiley & Sons (1989) and Sambrook *et al.*, Molecular Cloning A Laboratory Manual, 2nd Ed., Vols. 1-3, Cold Springs Harbor Press (1989). For example, hybridization and washing conditions may be selected to detect substantially perfect matches. They may also be selected to allow discrimination of perfect matches and one base pair mismatches. They may also be selected to permit the detection of large amounts of mismatches. As an example, the wash may be performed at the highest stringency that produces results and that provides a signal intensity greater than approximately 10% of the background intensity.

In hybridization between array-immobilized polynucleotide probes and the mixture of the target nucleic acids and control nucleic acids for detecting sequence variations, the target nucleic acids are typically tagged with a detectable label. Control nucleic acids which contain the reference sequence are also tagged with a detectable label. The labels for the target nucleic acids and the control nucleic acids are different. For example, Cy3 (green) may be used for control nucleic acid labeling and Cy5 (red) may be used for target nucleic acid labeling. Preferably, the control and target nucleic acids are mixed prior to or during the hybridization assay.

V. Determining the differences in hybridization between probes and target nucleic acids

After the initial set of polynucleotide probes is immobilized on an array and hybridized to the target nucleic acid, the hybridization intensities indicating the hybridization extent between the target nucleic acid and polynucleotide probes are determined and compared. The differences in hybridization intensities are evaluated. One of skilled in the art will appreciate that methods for evaluating the hybridization results vary with the nature of probes, sequence variations, gene expressions, and labeling methods. For example, quantification of the fluorescence intensity is accomplished by measuring probe signal strength at locations where probes are

present. Comparison of the absolute intensity of array-immobilized polynucleotide probes hybridized to target nucleic acids with intensities produced by mismatch probes and/or control probes provides a measure of the sequence variations or the expression of the genes.

Quantification of the hybridization signal can be by any means known to one of skill in the art. For example, quantification may be achieved by the use of a confocal fluorescence microscope. The methods of measuring and analyzing hybridization intensities may be performed utilizing a computer. The computer program typically runs a software program that includes computer code for analyzing hybridization intensities measured. Signals may be evaluated by calculating the difference in hybridization signal intensity between each polynucleotide probe, its mismatch probes, and control probes. The differences can be evaluated for each sequence variation or each gene. Examples of quantification of hybridization signals are shown in U.S. Patent Nos. 5,733,729, 5,974,164, 6,066,454, and 6,171,793.

Background signals typically contribute to the observed hybridization intensity. The background signal intensity refers to hybridization signals resulting from non-specific binding, or other interactions, *e.g.*, between target nucleic acids and array surface. Background signals may also be produced by the array component itself. A background signal may be calculated for an array and/or for each sequence variation or each gene expression analysis. For example, background may be calculated as the average hybridization signal intensity for the lowest 5% to 10% of the probes in the array, or where a different background signal is calculated for each sequence variation or gene, for the lowest 5% to 10% of the probes for each sequence variation or gene. Background signal may also be calculated as the average hybridization signal intensity produced by hybridization to probes that are not complementary to any sequence found in the sample (*e.g.* probes directed to nucleic acids of the opposite sense or to genes not found in the sample). Background may also be calculated as the average signal intensity produced by regions of the array that lack any probes at all. Preferably the difference in hybridization signal intensity between each probe and its control probes is detectable, *e.g.* greater than about 10%, 20%, or 50% of the background signal intensity. In some instances, only those probes where difference between the probe and its control probes exceeds a threshold hybridization intensity (*e.g.* preferably greater than 10%, 20%, or 50% of the background signal intensity) are selected. Thus, only probes that show a strong signal

compared to their control probes are selected. In addition, methods for correcting the effect of cross-hybridization in a hybridization assay are disclosed in WO 00/03039.

The identity of each sequence variation or the expression level of each gene may be estimated using known methods in the art. If the target is present, the perfectly matched probes should have consistently higher hybridization intensity than the mismatched probes. In some cases, the highest intensity probe may be compared to the second highest intensity probe. The ratio of the intensities may be compared to a predetermined ratio cutoff. Of course, ratio cutoff may be adjusted to produce optimal results for a specific array and for a specific sequence variation or a gene profiling. In addition to comparing to mismatch probes, the hybridization intensity may be compared to other probes, such as normalization probes. For example, probe intensity of target nucleic acid may be compared to that of a known sequence. Any significant changes may indicate the presence or absence of a sequence variation or a gene expression level. Statistical method may also be used to analyze hybridization intensities in determining sequence variations or gene expression levels. For example, mismatch probe intensities may be averaged. Means and standard deviations may be calculated and used in determining sequence variations and profiling gene expressions. Complex data processing and comparative analysis may be found in EP 717 113 A2 and WO 97/10365, both incorporated herein by reference.

As another example, in the case where the control nucleic acids are labeled with dye Cy3 (green) and the target nucleic acids are labeled with dye Cy5 (red), if the sequence variation in the target and control nucleic acids is identical, the resulting color may be yellow due to the mixing of similar amounts of target and control nucleic acids. If the sequence variation in target nucleic acids is different from that in the control nucleic acids, a subset of polynucleotide probes may only hybridize perfectly to the control nucleic acids, thus resulting a color shift from probes that hybridize perfectly to both the control and target nucleic acids. Thus, any significant changes in fluorescent intensity may indicate the presence or absence of the sequence variation of the control nucleic acids in the target nucleic acids.

A diploid organism may be homozygous or heterozygous for a polymorphic form or for a mutation. There are four possible homozygotes (A/A, T/T, C/C, and G/G) and six possible heterozygotes (T/A, A/G, C/T, C/A, T/G, and C/G). When the polynucleotide probes are hybridized with a heterozygous sample, the patterns for the homozygous samples are superimposed. Thus, the probes show distinct and

characteristic hybridization patterns depending on which sequence variation is present and whether an individual is homozygous or heterozygous.

Quantifying transcription levels of multiple genes can be absolute or relative quantification. Absolute quantification may be accomplished by inclusion of known concentration of one or more target nucleic acids such as control nucleic acids or known amounts of the target nucleic acids to be detected. The relative quantification may be accomplished by comparison of hybridization signals between two or more genes, or between two or more treatments to quantify the changes in hybridization intensity.

VI. Optimizing hybridization performance of polynucleotide probes

Although sequence variations or gene expressions of a target nucleotide acid are estimated as well as possible from the hybridization pattern to the initial array design, in most cases, not all sequence variations or gene expressions can be determined simultaneously under a given set of condition. Ambiguities may arise from the initial set of probes due to non-specific binding, cross-hybridization, base probe composition effect, and other factors. In particular, in gene expression profiling, accurately profiling gene expression levels is based on numerical assessment of hybridization intensities of the target to the probes, thus making the optimal probe selection even more crucial. Additional set(s) of polynucleotide probes is then designed based on the hybridization analysis of the initial probe set. For example, new generations of probes may be designed to maximize the discrimination ratio between matches and mismatches or to balance the stability of mismatches.

A. Polynucleotide Sequence and Length:

One of the factors influencing hybridization performance of a polynucleotide probe is base composition. It is well known that sequences rich in G/C are more stable than sequences with lower G/C content. The solution melting temperature (Tm) of a polynucleotide, at which 50% of the polynucleotide is hybridized and 50% is not hybridized, is often used as a practical indicator of the hybridization strength of a polynucleotide probe of a given base composition. Methods for measuring Tm of a polynucleotide are well known in the art. See, *e.g.*, Cantor and Schimmel, *Biophycical Chemistry*, San Francisco, W.H. Freeman (1980), incorporated herein by reference. There are also many ways to calculate Tm using mathematical algorithm.

A widely used rule of thumb is two degree of increase in Tm by adding an A/T base pair and four degree of increase in Tm by adding a G/C base pair. This simple formula may be further modified to take into account of the ionic strength and solvent effect. For example, Tm may be calculated using the formula:

$$Tm=81.5 + 16.6 (\log Na^+) + 0.41 \times \% \text{ of G/C} - 600/n - 0.65 \times \% \text{ of formamide}$$

Where Na^+ is sodium concentration, n is length of polynucleotide.

A more reliable formula to calculate Tm is available based on the interactions between a particular base and its nearest neighbors, *i.e.*, the nearest-neighbor model. An enthalpy and entropy for each nearest neighbor combination of two adjacent base pairs (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT) have been established based on the extensive melting experiments using various polynucleotide sequences. Thermodynamic coefficients of nearest-neighbor models are available for DNA/DNA, DNA/RNA, and RNA/RNA hybridizations. Therefore, free energy of hybridization of two sequences at any temperature in solution may be calculated. See, *e.g.*, U.S. Patent No. 5,556,749, Hyndman, D., *et al., BioTechniques* 20(6):1090-1097 (1996), Mitsuhashi, M., *J. Clinical Laboratory Analysis* 10:277-284 (1996), Wetmur, J., *Critical Reviews in Biochemistry and Molecular Biology,* 26:227-259 (1991), Rychlik *et al., Nucleic Acids Res.* 17:8543-8551 (1989), and Rychlik *et al., Nucleic Acids Res.* 18:6409-6412 (1990), all incorporated herein by reference.

The hybridization behavior of immobilized polynucleotide probes on a solid support is different from that in solution. Therefore, a more empirical approach is necessary to predict and modulate hybridization behavior of array-immobilized polynucleotide probes. Additional melting temperature experiments on solid supports may be conducted to more accurately characterize the thermodynamics and kinetics of hybridization behaviors of polynucleotide probes on an array. See Cantor and Smith, *Genomics: the science and technology behind the human genome project*, John Wiley & Sons (1999). Despite the differences in solid phase and solution phase kinetic and thermodynamic hybridization profiles, many variables affecting melting temperatures for solution hybridization, such as the effects of length, temperature, ionic strength, and solvent, are applicable for hybridization on solid supports.

In one embodiment of the present invention, Tm or free energy of hybridization may be evaluated based on base compositions, polynucleotide length, ionic strength, and thermodynamic parameters. High G/C content polynucleotide probes with a few mismatches may exhibit more stable hybridization than AT-rich

polynucleotides without mismatches. Mismatches in the middle of the probe sequence are more consequential for hybridization than those at the 5' or 3' end. Shorter probe lengths may provide the maximum mismatch destabilization and result in the greater match to mismatch ratios. However, this advantage is partially offset by the wide range of Tm values for short probes, depending on their specific sequence composition. For example, probes with 17 nucleotides long with a single base difference may differ by 5°C in Tm. If an array with equal length polynucleotide probes is used, baseline hybridization may yield wide range of signal intensities due to wide range of Tm values.

One skilled in the art will appreciate that in order to increase or decrease the melting temperature of a probe, it may be desirable to add, delete or change one or more bases in the probes. In certain embodiments of the inventions, polynucleotide probes with similar solution melting temperatures may be selected. An array of a plurality of polynucleotide probes may be fabricated wherein the melting temperatures of polynucleotide probes differ by no more than about 15 °C, more preferably by no more than about 10 °C, and still more preferably by no more than about 5 °C. The melting temperatures of polynucleotide probes immobilized on the array may also be within 10, 9, 8, 7, 6, or 5 °C of the average melting temperature. Typically, the better discrimination between matches and mismatches may be obtained at or slightly above the average Tm of polynucleotide probes. Alternatively, one may pre-select a hybridization temperature and then design polynucleotide probes that are within 5, 10, or 15 °C of the pre-selected hybridization temperature. The length of a polynucleotide probe may be changed by less than about 10, 5, or 2 nucleotides.

Consideration of secondary structure may also play a role in evaluating hybridization performance of polynucleotide probes, especially when high hybridization temperature to denature secondary structures may not be applied. If polynucleotides form secondary structure such as hairpins or triple helixes, intramolecular hybridization within polynucleotides may be energetically and kinetically favorable and they may not be available for hybridization to the target sequences. See Mitsuhashi, M., *J. Clinical Laboratory Analysis*, *supra*.

In order to design polynucleotides that are less likely to form secondary structures, one may calculate the free energies of secondary structure formation of all

candidate polynucleotide probes, based on the nearest-neighbor coefficients. Typically, polynucleotides having larger negative free energy form more stable hairpins, whereas polynucleotides having positive values or smaller negative values are less likely to form hairpins. One may select optimal polynuceotide probes based on the secondary structure energy values of the polynucleotide probes. There are also commercial software programs to predict the formation of secondary structure. In some instances, one may analyze the location of secondary structures by visual inspection. For example, palindromic sequences are known to readily form hairpin loops. If polynucleotide probes contain long stretches of CT or AG rich region, such an area may bind to double-stranded hybridization complex to form a triple helix structure.

In some instances, the presence of frequently appearing short subsequences may also be a factor for designing optimal polynucleotide sequences. For example, if polynucleotides contain a poly T or poly A stretch, such polynucleotides may cross-hybridize to poly(A)-mRNA or cDNA. If polynucleotides contain TATA-like sequences, such polynucleotides may bind to the promoter region of various genes.

A wide range of probe length may be used. Longer probes do not necessarily improve their sensitivity, because long probes usually exhibit higher Tm than that of actual assay conditions, allowing more mismatches. Although shorter probes increase the chances of nonspecific appearance of such sequences in the target sequences, they may exhibit a much higher penalty on mismatches. Therefore, one may design optimal probes based on their hybridization performance, instead of the length of the probes. In preferred embodiments of the present invention, the length of polynucleotide probes ranges from about 10 to about 100 nucleotides, preferably from about 10 to about 50 nucleotides.

In some instances, a combination of theoretical Tm balancing and empirical length adjustment may be employed. A probe set may be designed to have a common Tm, which provides uniform baseline hybridization signals from perfectly complementary probes. Then, within the probe group for each variation or each gene, probes may be shortened to maximize mismatch discrimination relative to the exact complement probe sequences. The resulting polynucleotide probe set may have uneven nucleotide lengths, but have more balanced Tm range. The length of polynucleotide probes may differ by about 10, 5, or 2 nucleotides while the melting temperatures of the probes may differ by no more than about 15, 10, or 5 °C.

B. Polynucleotide analogs

An alternative approach to even out base composition effects comprises the modification of one or more natural deoxynucleosides (or polynucleotide analogs) which forms a base pair whose stability is very close to that of the other pair. Polynucleotide analogs include base and sugar phosphate backbone analogs. An example of using polynucleotide analogs is shown in U.S. Patent 6,156,601.

Any base analogs that induce a decrease in stability of the three G/C hydrogen bonds or an increase in stability of the two A/T hydrogen bonds may be used. For example, one can substitute 2,6-diamino purine for A, which gives 2-NH_2A/T base pair having a stability similar to that of the G/C base pair. One may also select C derivatives, in which one hydrogen of the exocyclic amino group at position 4 is substituted by an alkyl group such as methyl, ethyl, *n*-propyl, allyl or propargyl groups. For example, a $G^{4Et}C$ base pair has stability similar to that of the A/T base pair. Typically, it may be easier to find a modified G/C base pair whose stability is similar to that of an A/T natural base pair than to design a modified A/T base pair whose stability is close to that of a G/C natural base pair. In addition, preparation of polynucleotides containing C analogs may be simpler than that of polynucleotides built with G analogs and modification of only one base pair rather than both may simplify the preparation of polynucleotides containing one or several modified nucleosides. Analogs that increase base stacking energy, such as pyrimidines with a halogen at the C5-position (*e.g.* 5-bromoU, or 5-ChloroU), may also be used. One may also use the non-discriminatory base analogue, or universal base, such as 1- (2–deoxy-D-ribfuranosyl)–3-nitropyrrole. This class of analogue maximizes stacking while minimizing hydrogen-bonding interactions without sterically disrupting a hybridization complex. See Nguyen, H., *et al.*, *Nucleic Acids Research* 25(15) 3059-3065 (1997) and Nguyen, H., *et al.*, *Nucleic Acids Research* 26(18):4249-4258 (1998), both incorporated herein by reference.

The highly charged phosphodiesters in natural nucleic acid backbone may be replaced by neutral sugar phosphate backbone analogues. The polynucleotide probes with uncharged backbones may be more stable, as in these analogs, the electrostatic repulsion between nucleic acid strands is minimized. As an example, phosphotriesters in which the oxygen that is normally charged in natural nucleic acids is esterified with an alkyl group may be used.

Another class of backbone analogs is polypeptide nucleic acids (PNAs), in which a peptide backbone is used to replace the phosphodiester backbone. The stability of PNA-DNA duplex is essentially salt independent. Thus low salt may be used in hybridization procedures to suppress the interference caused by stable secondary structures in the target. PNAs are capable of forming sequence-specific duplexes that mimic the properties of double-strand DNA except that the complexes are completely uncharged. Furthermore, because the hybridization stability of PNA-DNA is higher than that of DNA-DNA, binding is more specific and single-base mismatches are more readily detectable. See, *e.g.*, Giesen, U. *et al., Nucleic Acids Research* 26(21):5004-5006 (1998), Good, L., *et al., Nature Biotechnology* 16:355-358 (1998), and Nielsen, P., *Current Opinion in Biotechnology* 10:71-75 (1999), all incorporated herein by reference.

Another option to modulate the hybridization performance of polynucleotide probes is the replacement of naturally occurring nucleic acids have 3'-5' phosphodiester linkage. Polyribonucleotides with 2'-5' linkage which give complexes with lower melting temperature than duplexes formed by 3'-5' polynucleotides with the same sequence may be employed. See Kierzek, R., *et al., Nucleic Acids Research* 20(7):1685-1690 (1992), incorporated herein by reference.

Another method for optimizing hybridization performance is using polynucleotides containing C-7 propyne analogs of 7-deaza-2'-deoxyguanosine and 7-deaza-2'-deoxyadenosine (Buhr *et al., Nucleic Acids Res.* 24:2974-2980 (1996), incorporated herein by reference) or C-5 propyne pyrimidines (Wagner *et al., Science* 260:1510-3 (1993), incorporated herein by reference). These analogs may be particular useful in gene expression analysis.

C. Hybridization environment

Hybridization performance of polynucleotide is also dependent on the hybridization environment, for example, the concentrations of ions and nonaquous solvents. The hybridization performance of polynucleotide probes may be modulated by changing the dielectric constant and ionic strength of the hybridization environment. Salt concentrations, such as Na, Li, and Mg, may have an important influence on hybridization performance of polynucleotide probes.

Reagents that reduce the base composition dependence of hybridization performance may be used to alter the hybridization environment of array-immobilized

polynucleotide probes. For example, high concentrations of tetramethylammonium salts (TMAC), N,N,N,-trimethylglycine (Betain) may be added to target nucleic acid mixture. At suitable concentrations typically at multimolar concentrations, these reagents may equalize the Tm of polynucleotides that are pure A/T and those that are pure G/C and thus increase the discrimination between perfect matches and mismatches. See, Von Hipppel *et al.*, *Biochemistry*, 3:137-144 (1993) and U.S. Patent No. 6,045,996, incorporated herein by reference.

Denaturing reagents that lower the melting temperature of double stranded nucleic acids by interfering with hydrogen bonding between bases may also be used. Denaturing agents, which may be used in hybridization buffers at suitable concentrations (*e.g.* at multimolar concentrations), include formamide, formaldehyde, DMSO ("dimethylsulfoxide"), tetraethyl acetate, urea, GuSCN, and glycerol, among others.

Chaotropic salts that disrupt van der Waal's attractions between atoms in nucleic acid molecules may also be used. Chaotropic salts, which may be used in hybridization buffers at suitable concentrations (*e.g.* at multimolar concentrations), include, for example, sodium trifluoroacetate, sodium tricholoroacetate, sodium perchlorate, guanidine thiocyanate, and potassium thiocyanate, among others. See, Van Ness, J., *et al.*, *Nucleic Acids Research* 19(19):5143-5151 (1991), incorporated herein by reference.

Renaturation accelerants that increase the speed of renaturation of nucleic acids may also be used. They generally have relatively unstructured polymeric domains that weakly associate with nucleic acid molecules. Accelerants include cationic detergents such as, CTAB ("cetyltrimethylammonium bromide") and DTAB ("dodecyl trimethylammonium bromide"), and, heterogenous nuclear ribonucleoprotein ("hnRP") A1, polylysine, spermine, spermidine, single stranded binding protein ("SSB"), phage T4 gene 32 protein and a mixture of ammonium acetate and ethanol, among others. See, Pontius, B., *et al.*, *Proc. Natl. Acad. Sci. USA* 88:82373-8241 (1991), incorporated herein by reference.

One of skill in the art would appreciate that there are many other ways to modulate the hybridization performance of polynucleotides by changing the hybridization environment of polynucleotide probes. One method is changing the length of spacer that tethers polynucleotide probe to the array surface. It has been demonstrated that steric factors are important in increasing the efficiency of

hybridization between polynucleotide probes and target nucleic acids. See, Southern *et al.*, *Nucleic Acids Research*, 20(7):1679-1684 (1992), incorporated herein by reference. Methods for reducing non-specific binding to an array by surface modifications and probe modifications are described in WO 99/54509, incorporated herein by reference.

An alternative approach for enhancing the discrimination between matched and mismatches is applying electric current to polynucleotide probes which destabilize mismatches relative to matches. See, *e.g.*, U.S. Patent No. 5,929,208.

In some instances, the local concentration of polynucleotide probes or the concentration of target nucleic acids may be varied to allow maximum discrimination between matches and mismatches. In some instances, local concentrations of polynucleotide probes may be higher than target nucleic acids. Such high local DNA probe concentrations may generate high local charge densities and promote the undesirable association of probes that may interfere with target binding. High local probe concentration may also permit the simultaneous binding of target molecules to multiple probes, and may sterically prohibit access of target to the probes. If polynucleotide probes are at lower concentrations compared with the target sequence, the kinetics and thermodynamics of the hybridization may also be affected. See, Cantor and Smith, *supra*.

VII. Iterative design of polynucleotide probes.

Upon the redesign of the initial probe set, a second set of polynucleotide probes is immobilized on an array. The target nucleic acid is then hybridized with the second set of polynucleotide probes. Each sequence variation or gene expression is reestimated from the resulting hybridization pattern. Further cycles of array design and hybridization pattern analysis can be performed in an iterative fashion, if desired, until all sequence variations or gene expressions are determined under a single set of conditions.

VIII. Definitions

As used herein, the terms "polynucleotide" and "nucleic acid" refer to naturally occurring polynucleotides, *e.g.* DNA or RNA. This term also refers to analogs of naturally occurring polynucleotides. The polynucleotide may be double stranded or single stranded. The polynucleotides may be labeled with radiolabels,

fluorescent labels, enzymatic labels, proteins, haptens, antibodies, sequence tags. A target nucleic acid may include a control nucleic acid, although they may be labeled differently.

As used herein, the term "sequence variation" refers to a mutation or a polymorphic form. A polynucleotide variation may range from a single nucleotide variation to the insertion, modification, or deletion of more than one nucleotide. A sequence variation may be located at the exon, intron, or regulatory region of a gene. Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A biallelic polymorphism has two forms. A triallelic polymorphism has three forms. A polymorphic site is the locus at which sequence divergence occurs. Diploid organisms may be homozygous or heterozygous for allelic forms. Polymorphic sites have at least two alleles, each occurring at frequency of greater than 1% of a selected population. A mutation may occur at frequency of less than 1% of a selected population. Polymorphic sites also include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements. The first identified allelic form may be arbitrarily designated as the reference sequence and other allelic forms may be designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the wild type form (or the consensus sequence), which is frequently used as the reference sequence.

EXAMPLES OF THE PREFERRED EMBODIMENTS

The following examples further illustrate the present invention. These examples are intended merely to be illustrative of the present invention and are not to be construed as being limiting.

EXAMPLE 1

Polymorphisms, alleles, and phenotypes of the NAT2 Gene

N-acetyltransferase 2 (NAT2) is a polymorphic N-acetylation enzyme that detoxifies hydrazine and arylamine drugs and is expressed in the liver. The NAT2 coding region spans 872 base pairs (Genbank Accession No. NM-000015). The PCR product is approximately 1276 base pairs.

Polymorphisms in the NAT2 gene cause the fast and slow N-acetylation phenotypes implicated in the action and toxicity of amine containing drugs. In addition, NAT2 acetylation phenotype is associated with susceptibility to colorectal and bladder cancers. Table 1 summarizes the seven common single nucleotide polymorphisms (SNPs) found in this gene (G191A, C282T, T341C, C481T, G590A, A803G, and G857A) and defines the nine most common alleles (*4 being the wild type allele) along with their associated phenotypes and population frequencies. See Grant *et al.*, *Mutat. Res.* 376:61-70 (1997) and Spielberg *et al.*, *J. Pharmacokint. Biopharm.* 24:509-519 (1996). Each of the seven polymorphisms is a marker for more than one NAT2 allele and each variant allele is defined by two or three SNP substitutions. NAT2 provides a clearly defined, low complexity model system for developing a hybridization based genotyping assay. Typically, homozygous or heterozygous genotypes are made at each polymorphic site before probable allele assignments can be made. In general, individuals who are homozygous for any combination of the slow acetylator alleles are slow acetylators, where rapid acetylators are homozygous or heterozygous for wild-type NAT2 allele. It has been suggested that slow acetylators may be at increased risk for developing bladder, larynx and hepatocellular carcinomas, whereas rapid acetylator may be at risk to develop colorectal cancer. The frequency of the slow acetylator phenotype varies among ethnic groups and is roughly 50%-60% in Caucasian populations. See Grant, D., *et al*, *Mutation Research* 376:61-70 (1997) and Lin, H., *et al*, *Pharmacogenetics* 4:125-134 (1994). Polynucleotide array can be used to determine whether a target nucleic acid sequence has one or more nucleotides identical to or different from a specific reference sequence.

Table 1. Polymorphisms, alleles, and phenotypes of the NAT2 gene

po lym. allele	G191A	C282T	T341C	C481T	G590A	A803G	G857A	Phen.	Freq.
*4	G	C	T	C	G	A	G	Rapid	23.40
*5A	G	C	C	T	G	A	G	Slow	2.50
*5B	G	C	C	T	G	G	G	Slow	40.90
*5C	G	C	C	T	G	G	G	Slow	2.60
*6A	G	T	T	C	A	A	G	Slow	28.40
*7B	G	T	T	C	G	A	A	Slow	2.10
*12A	G	C	T	C	G	G	G	Rapid	0.10
*14A	A	C	T	C	G	A	G	Slow	rare
*14B	A	T	T	C	G	A	G	Slow	0.10
Amino Acid	R → Q	None	I → T	None	R → Q	K → R	G → E		

EXAMPLE 2

Preparation of array-immobilized polynucleotide probes

Surface tension array synthesis was a two step process, substrate surface preparation followed by *in situ* polynucleotide synthesis. Substrate preparation began with glass cleaning in detergent, then base and acid (2% Micro 90, 10% NaOH and 10% H_2SO_4) followed by spin coating with a layer of Microposit 1818 photoresist (Shipley, Marlboro, MA) that was soft baking at 90°C for 30 min. The photoresist was then patterned with UV light at 60 mWatts/cm^2 using a mask that defines the desired size and distribution of the array features. The exposed photoresist was developed by immersion in Microposit 351 Developer (Shipley, Marlboro, MA) followed by curing at 120°C for 20 minutes. Substrates were then immersed in 1% solution of tridecafluoro-1,1,2,2-tetrahydrooctyl)-1-trichlorosilane (United Chemical Technology, Bristol, PA) in dry toluene to generate a hydrophobic silane layer surrounding the array features which were still protected by photoresist. The fluorosilane was cured at 90°C for 30 min then treated with acetone to remove the remaining photoresist. The exposed feature sites were coated with 1% 4-aminobutyldimethylmethoxysilane (United Chemical Technology, Bristol, PA) then cured for 30 min at 105°C. Finally these sites were coupled with a linker molecule that will support subsequent polynucleotide synthesis.

These surface tension patterned substrates were aligned on a chuck mounted on the X-Y stage of a robotic array synthesizer where piezoelectric nozzles (Microfab Technologies, Plano, TX) were used to deliver solutions of activated standard H-phosphonate amidites (Froehler *et al.*, *Nucleic Acids Res.* 14:5339-5407 (1986)). The piezoelectric jets were run at 6.67 kHz using a two-step waveform, which fires individual droplets of approximately 50 picoliters. Washing, deblocking, capping, and oxidizing reagents were delivered by bulk flooding the reagent onto the substrate surface and spinning the chuck mount to remove excess reagents between reactions. The substrate surface was environmentally protected throughout the synthesis by a blanket of dry N_2 gas. Localizing and metering amidite delivery was mediated by a computer command file that directed delivery of the four amidites during each pass of the piezoelectric nozzle bank so a predetermined polynucleotide was synthesized at each array coordinate. Array design iterations were accomplished by altering this synthesis command file.

Piezoelectric printed polynucleotide synthesis was performed using the following reagents (Glen Research, Sterling, VA): phosphoramidites: pac-dA-CE phosphoramidite, Ac-dC-CE phosphoramidite, iPr-pac-dG-CE phosphoramidite, dT CE phosphoramidite (all at 0.1M); activator: 5-ethylthio tetrazole (0.45M). Amidites and activator solutions were premixed, 1:1:v/v, in a 90% adiponitrile (Aldrich, Milwaukee, WI): 10% acetonitrile solution prior to synthesis. Ancillary reagents were oxidizer (0.1M iodine in THF/pyridine/water), Cap mix A (THF/2,6-lutidine/acetic anhydride), Cap mix B (10% 1-methylimidazole/THF), and 3% TCA in DCM.

EXAMPLE 3

Target nucleic acids preparation, labeling and hybridization conditions

Hybridization target nucleic acids were prepared using PCR primers (5'-GTCACACGAGGAAATCAAATGC-3') (Seq. ID. No. 1) and 5'-GTTTTCTAGCATGAATCACTCTGC-3') (Seq. ID. No. 2) that amplify a 1.2 kb fragment from genomic DNA containing all 872 coding nucleotides in the single NAT2 exon as well as 5' and 3' non-coding sequences (Cascorbi *et al., Am. J. of Human Genetics* 57:581-591 (1995)). The PCR product was chromatographically purified, nicked with DNase to generate random fragments of about 50-100 nucleotides and end labeled in a TdT reaction with biotin-ddATP. This product was hybridized to microarrays for a minimum of two hours in 0.5M LiCl, 10mM Tris-HCl, pH 8.0, 0.005% sodium lauroyl sarcosein at 42 °C and washed in the same buffer without probe for 10 minutes at room temperature. Following washing, the hybridized, biotin-labeled targets were stained with a CY3-streptavidin conjugate (NEN-DuPont) covered with a microscope slide coverslip and imaged using the GenePix 4000 scanner (Axon Instruments, Foster City, CA).

Hybridization performance was analyzed by comparing intensities at intended complementary probe sites to each other and to known single and double mismatched probes. An ideal result is when perfect complements have high intensity signals that are essentially equivalent to each other and maximum discrimination ratios against mismatch probes.

EXAMPLE 4

Characterized hybridization samples

To evaluate different array designs and probe sets for their performance in discriminating among NAT2 genotypes, an anonymous set of genomic DNA samples with known NAT2 genotypes were obtained. These samples, which included a *4 homozygote and samples that collectively represented each of the seven common polymorphisms as heterozygotes, were sequenced to confirm the genotypes. Fluorescently labeled PCR products generated from this set of primary genomic samples were used in all of the array optimization studies.

EXAMPLE 5

Iterative array designs

In the first stage of the probe design, all probes on the array were of a single length. In particular, length of 17 nucleotides was chosen. Twenty polynucleotide probes were selected for the coding strand and 20 for the non-coding strand, giving a total of 40 probes for each polymorphism. Probes for both the coding strand and noncoding strand were designed such that the polymorphism site was at the center in each probe. For each polymorphic site, a full set of polynucleotide probes for the coding strand (20 total) includes one set of four 17-mers having A, C, G or T substituted at the center polymorphic site (4), another two sets of four 17-mers having A, C, G or T substituted at one nucleotide 5' to the center polymorphic site with either A or G as the reference sequence (8), another two sets of four 17-mers having A, C, G or T substituted at one nucleotide 3' to the center polymorphic site with either A or G as the reference sequence (8). A full set of polynucleotide probes for the non-coding strand (20) is constructed similarly. The cumulative result is a related set of probes perfectly complementary to each known polymorphism as well as a set of single and double nucleotide-mismatched control probes. For example, an initial set of polynucleotide probes for the coding strand detecting the G191A SNP is shown below:

Reference Coding sequence 5'-AGAAGAAACCCGGGTGGGTG-3' (Seq. ID. No. 3)

A substituted at the polymorphic site 3'-TTCTTTGGACCCACCC-5' (Seq. ID. No. 4)

C substituted at the polymorphic site	3'-TTCTTTGGCCCCACCC-5' (Seq. ID. No. 5)
G substituted at the polymorphic site	3'-TTCTTTGGGCCCACCC-5' (Seq. ID. No. 6)
T substituted at the polymorphic site	3'-TTCTTTGGTCCCACCC-5' (Seq. ID. No. 7)
A substituted at 3' of the polymorphic site with A	3'-TTCTTTGGTACCACCC-5' (Seq. ID. No. 8)
C substituted at 3' of the polymorphic site with A	3'-TTCTTTGGTCCCACCC-5' (Seq. ID. No. 9)
G substituted at 3' of the polymorphic site with A	3'-TTCTTTGGTGCCACCC-5' (Seq. ID. No. 10)
T substituted at 3' of the polymorphic site with A	3'-TTCTTTGGTTCCACCC-5' (Seq. ID. No. 11)
A substituted at 3' of the polymorphic site with G	3'-TTCTTTGGCACCACCC-5' (Seq. ID. No. 12)
C substituted at 3' of the polymorphic site with G	3'-TTCTTTGGCCCCACCC-5' (Seq. ID. No. 13)
G substituted at 3' of the polymorphic site with G	3'-TTCTTTGGCGCCACCC-5' (Seq. ID. No. 14)
T substituted at 3' of the polymorphic site with G	3'-TTCTTTGGCTCCACCC-5' (Seq. ID. No. 15)
A substituted at 5' of the polymorphic site with A	3'-TTCTTTGATCCCACCC-5' (Seq. ID. No. 16)
C substituted at 5' of the polymorphic site with A	3'-TTCTTTGCTCCCACCC-5' (Seq. ID. No. 17)
G substituted at 5' of the polymorphic site with A	3'-TTCTTTGGTCCCACCC-5' (Seq. ID. No. 18)
T substituted at 5' of the polymorphic site with A	3'-TTCTTTGTTCCCACCC-5' (Seq. ID. No. 19)
A substituted at 5' of the polymorphic site with G	3'-TTCTTTGACCCCACCC-5' (Seq. ID. No. 20)

C substituted at 5' of the polymorphic site with G 3'-TTCTTTGCCCCCACCC-5' (Seq. ID. No. 21)

G substituted at 5' of the polymorphic site with G 3'-TTCTTTGGCCCCACCC-5' (Seq. ID. No. 22)

T substituted at 5' of the polymorphic site with G 3'-TTCTTTGTCCCCACCC-5' (Seq. ID. No. 23)

The full Tm range estimated for the perfect complement probes of 17-mers was 14.5°C and for mismatch controls, 19.3°C. Fluorescence intensity of at least three times background was observed for only 57% of the probe sets under the assay conditions used. For these probes, the average homozygous discrimination ratio, calculated as the average fluorescence signal from exact complement probes for each variant form divided the average fluorescence signal from mismatch controls, was 4.31.

Modification made to these probes for the second array iteration included lengthening probes that did not give hybridization signal intensities greater than three times background and shortening poorly discriminating probes. Probe length for version 2 ranged from 16 to 20 nucleotides. Mismatch positions were placed as close as possible to the center of probe sequences. This resulted in more total probe sets giving fluorescence signals above the cutoff value after hybridization and a good heterozygote discrimination ratio (exact matches of one variant /exact matches of the second variant) of 1.04. However, there was no significant improvement in the homozygote discrimination ratio.

Arrays with Tm balanced probe sets

The third array iteration was a complete array redesign based on calculated thermal melting points for the probe-target duplexes. The targeted Tm for every probe in the set was 63°C. Algorithm Tm = 81.5 + (100 * 0.41 * percent GC) – (675/length) was used to calculate solution Tm. Probe lengths ranged from 15 to 23 nucleotides and perfect match Tm ranged from 61.1 to 64.5 °C.

The polymorphism site was generally centered in the probes and the probe length was allowed to vary as needed to match the targeted Tm value. Hybridization to this array resulted in positive fluorescence signals for 100% of the probe sets but also resulted in substantial reduction of the global array homozygote discrimination ratio to 3.26. This reduction in sequence discrimination most likely reflects the strong

selection criteria globally applied to the probes for similar hybridization stability. However, in order to arrive at an array design capable of detecting specific genotypes by hybridization, maximizing relative fluorescence intensity for exact complement probes relative to negative mismatched controls is preferred. Therefore this selection criteria was emphasized during subsequent design modifications to further optimize the array genotyping performance. Figure 1 illustrates the global homozygote and heterozygote discrimination ratio values for each NAT2 genotyping array design iteration while Table 2 summarizes the performance characteristics for each array version. Two final design optimization cycles applied to the Tm selected probe design resulted in genotyping array version six which has a global homozygote discrimination ratio of 6.6 and an average heterozygote discrimination ratio of 1.0.

Table 2. Summary of the performance characteristics for six array versions.

Array Version / Probe Characteristic	Version 1	Version 2	Version 3	Version 4	Version 5	Version 6
Polymorphism 1 Length Range	17	16-20	15-23	15-23	15-23	14-23
Polymorphism 1 Average Length	17	19	18.38	18.44	18.44	18.63
Polymorphism 1 Tm Range	14.5	16.1	3.4	3.4	3.4	6.3
Polymorphism 1 Average Tm	60.18	63.15	63.10	63.16	63.16	63.27
Polymorphism 1 Average %GC	45	45	46	46	46	45
Polymorphism 2 Length Range	17	16-20	15-22	15-22	15-22	15-22
Polymorphism 2 Average Length	17	19	19.05	19.07	18.85	18.76
Polymorphism 2 Tm Range	14.5	18.76	5.1	4.2	6.3	7.3
Polymorphism 2 Average Tm	59.37	63.62	63.73	63.78	63.33	63.02
Polymorphism 2 Average %GC	43	44	44	44	44	44
Mismatch Length Range	17	16-20	15-23	14-23	13-23	13-23
Mismatch Average Length	17	18	18.54	18.07	17.30	17.30
Mismatch Tm Range	19.3	21.2	6.8	6.3	6.80	6.80
Mismatch Average Tm	60.28	62.97	63.4	62.53	60.78	60.79
Mismatch Average %GC	45	45	46	46	46	46
Average Homozygote Discrim. Ratio	4.31	ND	3.26	6.64	5.93	6.60
Average Heterozygote Discrim. Ratio	ND	1.04	1.17	ND	0.94	1.00

EXAMPLE 6

Optimized probes for genotyping the T341C polymorphism

A detailed example of probe set optimization for a specific polymorphism is shown in Figure 2A-2B. The first panel (Figure 2A) shows a typical hybridization to the constant length probe set in the first array design. Cross hybridization to mismatch control probes was clearly evident. Coding strand probes have a homozygote discrimination ratio of 2.3 and non-coding strand probes a ratio of 4.7.

The average probe Tm is 68°C. The second panel (Figure 2B) shows a typical hybridization to the third array design iteration which has Tm matched probes targeting 64°C. Although the global discrimination ratio was poorer for this array than for the first array iteration, discrimination ratios for the T341C polymorphism improved substantially over the first array iteration. Nevertheless there is still significant cross hybridization to the negative control probes. In Figure 3, hybridization results are shown from using the fully optimized array to genotype two patient samples, one that is heterozygous for the T341C polymorphism and one that is homozygous for "T" at position 341. The average calculated Tm for this final probe set is 61°C and the homozygote discrimination ratios are 6.9 for the coding strand probes and 10.7 for the non-coding strand probes.

EXAMPLE 7

Patient sample genotyping results

The optimized NAT2 genotyping array, design iteration 6, was used to genotype seventeen genomic DNA samples from renal failure patients. These genotypes were done as part of a broader study undertaken to assess whether any association exists between the renal failure phenotype and NAT2 metabolic enzyme genotype. Table 3 shows microarray-based genotype assignments for each of the seventeen patient samples as well as the most probable allele asignments and their associated phenotype predictions. To confirm the array-based genotype assignments, all 872 coding nucleotides of the NAT2 gene were sequenced in each of the seventeen genomic samples. Perfect concordance was found between the microarray assigned genotypes and the Sanger sequence data.

Table 3. Microarray-based genotype assignments for seventeen patient samples and their associated phenotype predictions.

Sample	G191A	C282T	T341C	C481T	G590A	A803G	G857A	Allel s	Phenotype
230/RW	G	C/T	T	C	G/A	A	G	*4/*6A	Rapid/Slow
188/TFD	G	C/T	T	C	G	A	G/A	*4/*7B	Rapid/Slow
343/AB	G	C/T	T	C	G	A	G/A	*4/*7B	Rapid/Slow
449/FF	G	C/T	T/C	C/T	G/A	A	G	*5B/*12A	Rapid/Slow
623/DF	G	C	T/C	C/T	G	G	G	*5B/*4	Rapid/Slow
465/MM	G	C	T/C	C/T	G	G/A	G	*5B/*4	Rapid/Slow
30/SM	G	C	T/C	C/T	G	G/A	G	*5B/*4	Rapid/Slow
641/EVB	G	C	T/C	C/T	G	G/A	G	*5B/*4	Rapid/Slow
172/JM	G	C	T/C	C/T	G	G/A	G	*5B/*4	Rapid/Slow
147/BR	G	C	T/C	C/T	G	G/A	G	*5B/*4	Rapid/Slow
120/GR	G	C	T/C	C/T	G	G/A	G	*5B/*5B	Slow
544/WFR	G	C	C	T	G	G	G	*5B/*5B	Slow
173/AF	G	C	C	T	G	G	G	*5B/*5C	Slow
443/IMM	G	C	C	C/T	G	G	G	*5B/*6A	Slow
356/NH	G	C/T	T/C	C/T	G/A	G/A	G	*6A/*5A	Slow
305/AT	G	T	T	C	A	A	G	*6A/*6A	Slow
399/CH	G	T	T	C	A	A	G	*6A/*6A	Slow

Note: C/T or G/A indicates heterozygosity at that polymorphism

EXAMPLE 8

Optimization of probes for gene expression profiling.

As an example of probe iteration in gene expression profiling, optimization the beta actin gene (GenBank accession number AB004047) has been chosen and is shown in Examples 8-10. This form of actin is a constituent of the cytoskeleton of non-muscular cells. Because of its high abundancy, the β-actin gene is used frequently in research laboratories for normalization of mRNA or gene expression profiles.

Preparation of array-immobilized polynucleotide probes is similar to that described in Example 2. For the first array design, two probes were selected to monitor the expression of the actin gene. The starting nucleotide location of the probes were 335 and 600 with the sequences of 5'-GTACTAGACCCAGTAGAAGAGCGCCAACCGGAACCCCAAGTCCCC-3') (Seq. ID. No. 24), and 5'-ACAGTGCGTGCTAAAGGGCGAGCCGGCACCACCACTTCGACATCG-3' (Seq. ID. No. 25), respectively. In this array design, single length probes of 45 nucleotides were chosen. Dig-labeled cRNA was used as the target nucleic acid. Hybridization was carried out overnight in 1x MES buffer at 65 °C in the presence of 0.1 mg/ml herring sperm DNA and 0.5 mg/ml acetylated BSA for blocking nonspecific binding sites. The result is shown in Figure 4. Although probes 1 (starting at position 335)

and 2 (starting at position 600) were both 45 base pairs in length. Probe 1 produced a significantly less intense signal than probe 2.

EXAMPLE 9

Array design iteration.

The difference in hybridization intensity in Example 8 is likely the result of secondary structures and cross hybridization. In order to explore the hybridization behavior of probes starting at various different locations of the actin gene, 22 new probes were designed (Table 4 and Figure 5). From the hybridization intensities it can be easily observed that probe location indeed influences the hybridization of the target to the probes. The probes selected in Example 8 (black bars in Figure 5) are in areas of low hybridization signal (probe 335) and good hybridization signal (probe 600). These findings explain the observation made in Example 8 and demonstrate the importance of careful probe design.

Table 4. Probes sequences of the second array iteration in actin gene profiling.

Pos.	Seq.	Tm	SeqID
64	CTCCCCTTCTGCCGGGCTCCCCGTAGCAGCGGGCGCTT	102	26
121	CTTAGGAAGACTGGGTACGGGTGGTAGTGCGGGACCAC	97	27
301	CCCAAGTCCCCCCGGAGCCAGTCGTCGTGCCCCACGAG	100	28
312	CCAACCGGAACCCCAAGTCCCCCCGGAGCCAGTCGTCG	100	29
331	TAGACCCAGTAGAAGAGCGCCAACCGGAACCCCAAGTC	97	30
408	ATGCCGGTCTCCGCATGTCCCTGTCGTGTCGGACCTAC	96	31
439	GGCAGTGGCCTCAGGTAGTGCTACGGTCACCATGCCGG	98	32
440	GGGCAGTGGCCTCAGGTAGTGCTACGGTCACCATGCCG	98	33
445	CACTGGGGCAGTGGCCTCAGGTAGTGCTACGGTCACCA	97	34
478	TCCCGTATGGGGAGCATCTACCCGTGTCACACCCACTG	96	35
489	ACCGTACCCCCTCCCGTATGGGGAGCATCTACCCGTGT	97	36
495	CGTCCTACCGTACCCCCTCCCGTATGGGGAGCATCTAC	97	37
497	TGCGTCCTACCGTACCCCCTCCCGTATGGGGAGCATCT	97	38
515	GGCCGGTCGGTCCAGGTCTGCGTCCTACCGTACCCCCT	99	39
521	GTCCAGGGCCGGTCGGTCCAGGTCTGCGTCCTACCGTA	100	40
669	TCCTCGATCTTCGGCGGCACCGGTAGAGGACGAGCTTC	97	41
672	CCCTCCTCGATCTTCGGCGGCACCGGTAGAGGACGAGC	97	42
679	AAGAGGTCCCTCCTCGATCTTCGGCGGCACCGGTAGAG	97	43
769	GGGTCCTTCCTTCCAACCTTCTCTCGGAGTCCCGTCGC	97	44
775	AGGTACGGGTCCTTCCTTCCAACCTTCTCTCGGAGTCC	96	45
1025	GACCTTCCACCTGTCGCTCCGGTCCTACCTCGGCGGCT	98	46
1034	GGTGTAGACGACCTTCCACCTGTCGCTCCGGTCCTACC	98	47

Tm values are calculated using Tm=81.5+(100*0.41*%GC)-(675/n), where n is the number of nucleotides of the probe.

EXAMPLE 10

Polynucleotide probes with different lengths

The use of short polynucleotide probes for gene expression profiling has the advantage over long cDNA probes that cross hybridization to targets with high homologies in sequence can be prevented. On the other hand, using short probes may lead to the loss of target discrimination. Figure 6 shows the hybridization pattern for three different probe lengths selected at three different probe locations for the β-actin gene. The hybridization signals obtained indicate that the location and length of the probe determines the hybridization signal intensity. To investigate and ensure hybridization specificity, mismatches were introduced in the center of the probes. Three mismatches (3 MM) and five mismatches (5 MM) reduced the signal intensity to approximately 50% and 25%, respectively. This indicates that the probes are hybridizing specifically and that hybridization conditions are chosen appropriately.

EXAMPLE 11

Preparation of array-immobilized polynucleotide probes

Surface tension array synthesis is a two step process, substrate surface preparation followed by *in situ* polynucleotide synthesis. Substrate preparation begins with glass cleaning in detergent, then base and acid (2% Micro 90, 10% NaOH and 10% H_2SO_4) followed by spin coating with a layer of Microposit 1818 photoresist (Shipley, Marlboro, MA) that is soft baking at 90°C for 30 min. The photoresist is then patterned with UV light at 60 mWatts/cm^2 using a mask that defines the desired size and distribution of the array features. The exposed photoresist is developed by immersion in Microposit 351 Developer (Shipley, Marlboro, MA) followed by curing at 120°C for 20 minutes. Substrates are then immersed in 1% solution of tridecafluoro-1,1,2,2-tetrahydrooctyl)-1-trichlorosilane (United Chemical Technology, Bristol, PA) in dry toluene to generate a hydrophobic silane layer surrounding the array features which are still protected by photoresist. The fluorosilane is cured at 90°C for 30 min then treated with acetone to remove the remaining photoresist. The exposed feature sites are coated with 1% 4-aminobutyldimethylmethoxysilane (United Chemical Technology, Bristol, PA) then cured for 30 min at 105°C. Finally these sites are coupled with a linker molecule that will support subsequent polynucleotide synthesis. As shown in Figure 7, four

perfectly matched interrogation probes (5'-TCCAGGTAGT-3' (Seq. ID. 48), 5'-AGTGCGTATC-3' (Seq. ID. No. 49), 5'-GTAGCAGTAG-3' (Seq. ID. No. 50), and 5'-TCCAGTTCGT-3' (Seq. ID. No. 51) are designed for a reference sequence (5'-ACTACCTGGATACGCACTACTGCTACGAACTGGT-3' (Seq. ID. No. 52)). The interrogation probes can vary in length. The interrogation probes overlap each other at the 3' and 5' end. This overlap can be one or more base pairs long. Due to the loss of discrimination for potential mismatches occurring at either the 3' or the 5' end of the interrogation probes, the interrogation probes are overlapping, providing an easy additional control (the assumption is that the mismatch is only real when the mismatch signal is observed in both overlaping probes, compare Fig. 9).

EXAMPLE 12

Target nucleic acids preparation and two-color fluorescent analysis

Target nucleic acids are prepared using PCR primers that amplify a fragment from genomic DNA. The PCR product is chromatographically purified, nicked with DNase to generate random fragments of about 50-100 nucleotides and end labeled with dye Cy5. A control sequence containing Seq. ID. 52 is end labeled with dye Cy 3. The target and control nucleic acids are mixed (Figure 8). This mixture is hybridized to the array for about 3 hours in 0.5M LiCl, 10mM Tris-HCl, pH 8.0, 0.005% sodium lauroyl sarcosein at 42 °C and washed in the same buffer without probe for 10 minutes at room temperature. Following washing, the extent of hybridization between the control/target nucleic acid mixture and the immobilized polynucleotide probes are analyzed with an image scanner.

Figure 9 shows the results of two-color fluorescent analysis. In Panels A, B and C, the control nucleic acids which are labeled with dye Cy3; target nucleic acids which contain either one of the three sequence variations are labeled with dye Cy5. In Panel A, probes 1, 3, and 4 have perfect hybridization with both the control and target nucleic acids. However, probe 2 hybridizes perfectly only with the control nucleic acids and has a mismatch with the target nucleic acids at the sequence variation position (indicated by a star sign). The fluorescent intensity at probe 2 is different from that at probes 1, 3 and 4. If the target nucleic acids contain the identical sequence variation as the control sample (not shown in the figure), a different hybridization pattern results. In Panel B, probes 1 and 4 have perfect hybridization

with both the control and target nucleic acids. Probes 2 and 3 hybridize perfectly only with the control nucleic acids and have a mismatch with the target nucleic acids. The fluorescent intensities at probes 2 and 3 are different from those at probes 1 and 4. Panel C is similar to Panel A except that probe 3 shows a different hybridization result.

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. These variations may be applied without departing from the scope of the invention. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

All publications, patents, web sites are herein incorporated by reference in their entirety to the same extent as if each individual publication, patent or web site was specifically and individually indicated to be incorporated by reference in its entirety.

CLAIMS:

1. A method for simultaneously determining the presence or absence of two or more sequence variations in target nucleic acids, comprising the steps of:

(a) obtaining an array wherein polynucleotide probes are immobilized on said array;

(b) hybridizing said target nucleic acids to said polynucleotide probes under a pre-determined condition;

(c) determining differences in hybridization between target nucleic acids and said polynucleotide probes;

(d) changing the melting temperature of at least one said polynucleotide probe; and

(e) repeating steps (a)-(d), if necessary, until said differences in hybridization between said target nucleic acids and said polynucleotide probes simultaneously indicate the presence or absence of said two or more sequence variations in said target nucleic acids under said pre-determined condition.

2. The method according to claim 1 wherein step (d) is changing the length of at least one said polynucleotide probe.

3. The method according to claim 1 wherein step (d) is changing the sequence composition of at least one said polynucleotide probe.

4. The method according to claim 1 wherein step (d) is changing the hybridization environment of at least one said polynucleotide probe.

5. The method according to claim 1 wherein the melting temperature of step (d) is changed by no more than about 15 °C.

6. The method according to claim 1 wherein the melting temperature of step (d) is changed by no more than about 10 °C.

7. The method according to claim 1 wherein the melting temperature of step (d) is changed by no more than about 5 °C.

8. The method according to claim 2 wherein the length of said at least one polynucleotide probe is changed by no more than about 10 nucleotides.

9. The method according to claim 2 wherein the length of said at least one polynucleotide probe is changed by no more than about 5 nucleotides.

10. The method according to claim 2 wherein the length of said at least one polynucleotide probe is changed by no more than about 2 nucleotides.

11. The method according to claim 3 wherein said changing the sequence composition comprises using one or more polynucleotide analogs.

12. The method according to claim 1 wherein said sequence variations are polymorphic forms or mutations of a gene.

13. The method according to claim 1 wherein said polynucleotide probes are covalently linked to the surface of said array.

14. The method according to claim 1 wherein said polynucleotide probes are not covelently linked to the surface of said array.

15. The method according to claim 1 wherein said polynucleotide probes are synthesized *in situ*.

16. The method according to claim 1 wherein said polynucelotide probes are presynthesized prior to immobilization on the surface of said array.

17. The method according to claim 1 wherein said polynucleotide probes are separated by surface tension.

18. The method according to claim 1 wherein said immobilization is performed using an ink jet printing apparatus.

19. The method according to claim 1 wherein said immobilization is performed using a piezoelectric pump.

20. The method according to claim 1 wherein the lengths of said polynucleotide probes range from about 10 to 100 nucleotides.

21. A method for simultaneously monitoring the expression of two or more genes in target nucleic acids comprising the steps of:

(a) obtaining an array wherein polynucleotide probes are immobilized on said array;

(b) hybridizing said target nucleic acids to said polynucleotide probes under a pre-determined condition;

(c) determining differences in hybridization between target nucleic acids and said polynucleotide probes;

(d) changing the melting temperature of at least one said polynucleotide probe; and

(e) repeating steps (a)-(d), if necessary, until said differences in hybridization between said target nucleic acids and said polynucleotide probes simultaneously indicate levels of transcription of said two or more genes under said pre-determined condition.

22. The method according to claim 21 wherein step (d) is changing the length of at least one said polynucleotide probe.

23. The method according to claim 21 wherein step (d) is changing the sequence composition of at least one said polynucleotide probe.

24. The method according to claim 21 wherein step (d) is changing the hybridization environment of at least one said polynucleotide probe.

25. The method according to claim 21 wherein the melting temperature of step (d) is changed by no more than about 15 °C.

26. The method according to claim 21 wherein the melting temperature of step (d) is changed by no more than about 10 °C.

27. The method according to claim 21 wherein the melting temperature of step (d) is changed by no more than about 5 °C.

28. The method according to claim 22 wherein the length of said at least one polynucleotide probe is changed by no more than about 10 nucleotides.

29. The method according to claim 22 wherein the length of said at least one polynucleotide probe is changed by no more than about 5 nucleotides.

30. The method according to claim 22 wherein the length of said at least one polynucleotide probe is changed by no more than about 2 nucleotides.

31. The method according to claim 23 wherein said changing the sequence composition comprises using one or more polynucleotide analogs.

32. The method according to claim 21 wherein said target nucleic acids are a pool of RNAs.

33. The method according to claim 32 wherein said RNAs are *in vitro* transcribed from a pool of cDNAs.

34. The method according to claim 21 wherein said polynucleotide probes are covalently linked to the surface of said array.

35. The method according to claim 21 wherein said polynucleotide probes are not covalently linked to the surface of said array.

36. The method according to claim 21 wherein said polynucleotide probes are synthesized *in situ*.

37. The method according to claim 21 wherein said polynucleotide probes are presynthesized prior to immobilization on the surface of said array.

38. The method according to claim 21 wherein said polynucleotide probes are separated by surface tension.

39. The method according to claim 21 wherein said immobilization is performed using an ink jet printing apparatus.

40. The method according to claim 21 wherein said immobilization is performed using a piezoelectric pump.

41. The method according to claim 21 wherein the lengths of said polynucleotide probes range from about 10 to 100 nucleotides.

42. The method of claim 1 or 21 further comprising the step of estimating the melting temperatures of polynucleotide probes using a mathematical formula.

43. An array wherein the melting temperatures of polynucleotide probes immobilized on said array differ by no more than about 10 °C.

44. An array wherein the melting temperatures of polynucleotide probes immobilized on said array differ by no more than about 5 °C.

45. An array wherein the melting temperatures of polynucleotide probes immobilized on said array differ by no more than about 10 °C from the average melting temperature.

46. The array according to claims 43-45 wherein said polynucleotide probes differ in length by no more than 10 nucleotides.

47. The array according to claims 43-45 wherein said polynucleotide probes differ in length by no more than 10 nucleotides.

48. The array according to claims 43-45 wherein said polynucleotide probes differ in length by no more than 5 nucleotides.

49. A method for determining the presence or absence of a sequence variation in a target nucleic acid sequence comprising the steps of:

(a) immobilizing at least two polynucleotide probes on a solid support wherein at least one polynucletide probe spans the location of the sequence variation;

(b) attaching the target nucleic acid sequence with a first detectable label;

(c) attaching a control nucleic acid sequence with a second detectable label wherein the second detectable label is different than the first detectable label;

(d) contacting the immobilized polynucleotide probes with the mixture of the control nucleic acid sequence and the target nucleic acid sequence under hybridization conditions; and

(e) determining the presence or absence of the sequence variation in the target nucleic acid sequence based on the hybridization pattern differences of polynucleotide probes.

50. The method according to claim 49 wherein the sequence variation is a polymorphic form or mutation of a gene.

51. The method according to claim 49 wherein the polynucleotide probes are covalently linked to the surface of the solid support.

52. The method according to claim 49 wherein the polynucleotide probes are non-covalently attached to the surface of the solid support.

53. The method according to claim 49 wherein the polynucleotide probes are synthesized *in-situ*.

54. The method according to claim 49 wherein the polynucleotide probes are spotted on said solid support.

55. The method according to claim 49 wherein the solid support is a surface tension array.

56. The method according to claim 49 wherein said immobilization is performed using an ink jet printing apparatus.

57. The method according to claim 49 wherein the length of the polynucleotide probes range from about 6 to 100 nucleotides.

58. The method according to claim 49 wherein said at least two polynucleotide probes overlap by more than 1 base pair.

59. The method according to claim 49 wherein the density of polynucleotide probes on the surface of the solid support is between about 2-10,000 per cm^2.

60. The method according to claim 49 wherein said first and second labels are fluorescent labels.

61. The method according to claim 49wherein said solid support is glass.

62. The method according to claim 49 wherein the solid support is functionalized.

63. The method according to claim 49 wherein the size of each functionalized site is about 0.1×10^{-5} to 0.1 cm^2.

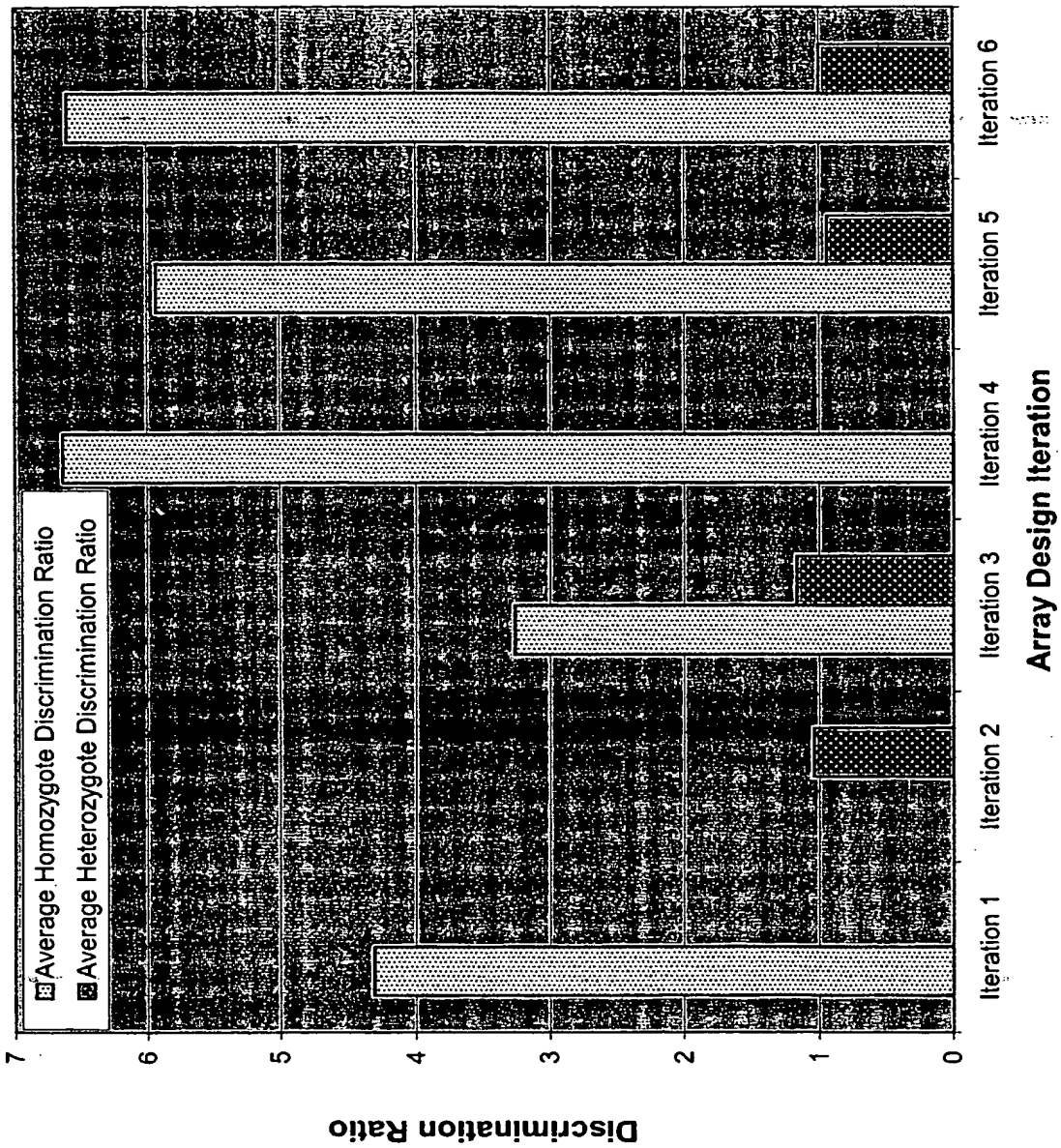
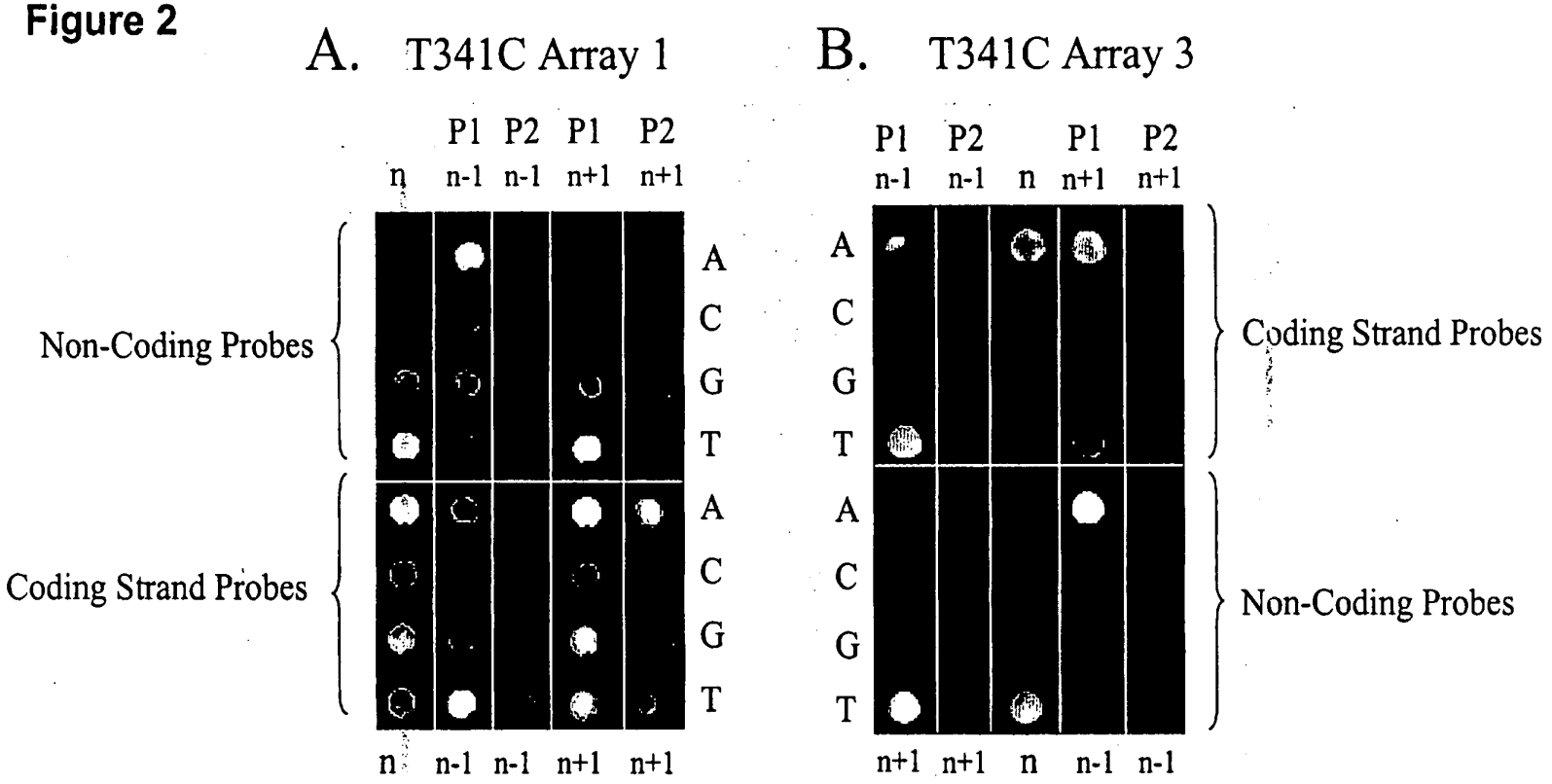


Figure 1

Figure 2

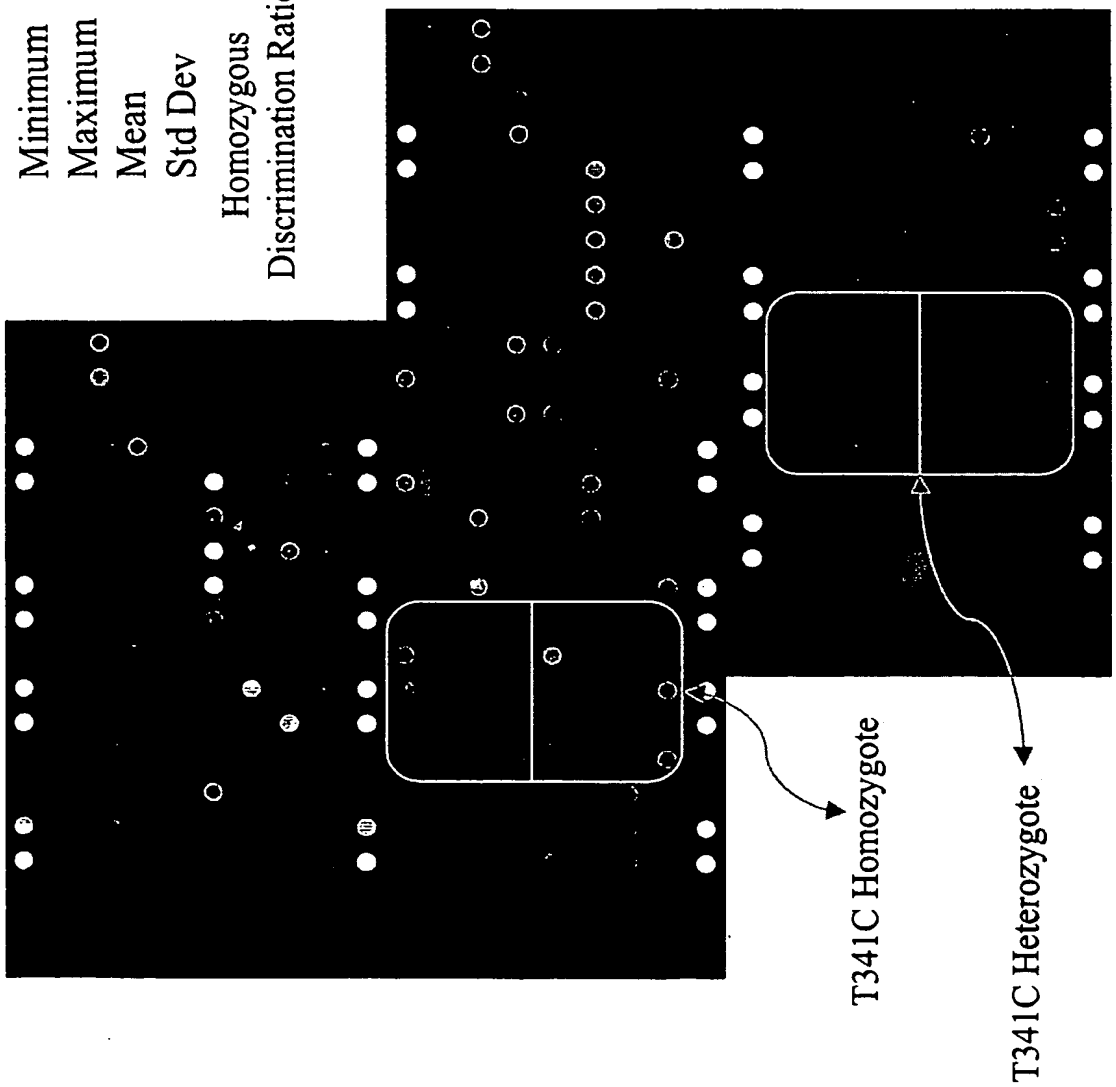


Design 1	L	Tm	%GC
Minimum	17	65.91	0.59
Maximum	17	70.74	0.71
Mean	17	68.08	0.64
Std Dev	0	1.71	0.04

Design 3	L	Tm	%GC
Minimum	15	62.38	0.56
Maximum	16	66.57	0.73
Mean	15.5	64.16	0.64
Std Dev	0.51	1.52	0.06

Homozygous Discrimination Ratios		
Coding	2.33	5.51
Non-Coding	4.67	6.86

Figure 3



Array Iteration 6

	L	Tm	%GC
Minimum	13	57.96	0.56
Maximum	16	63.83	0.77
Mean	15	61.07	0.64
Std Dev	0	1.94	0.06

Homozygous Discrimination Ratios: Coding 6.89; Non-Coding 10.7

Figure 4

Figure 5

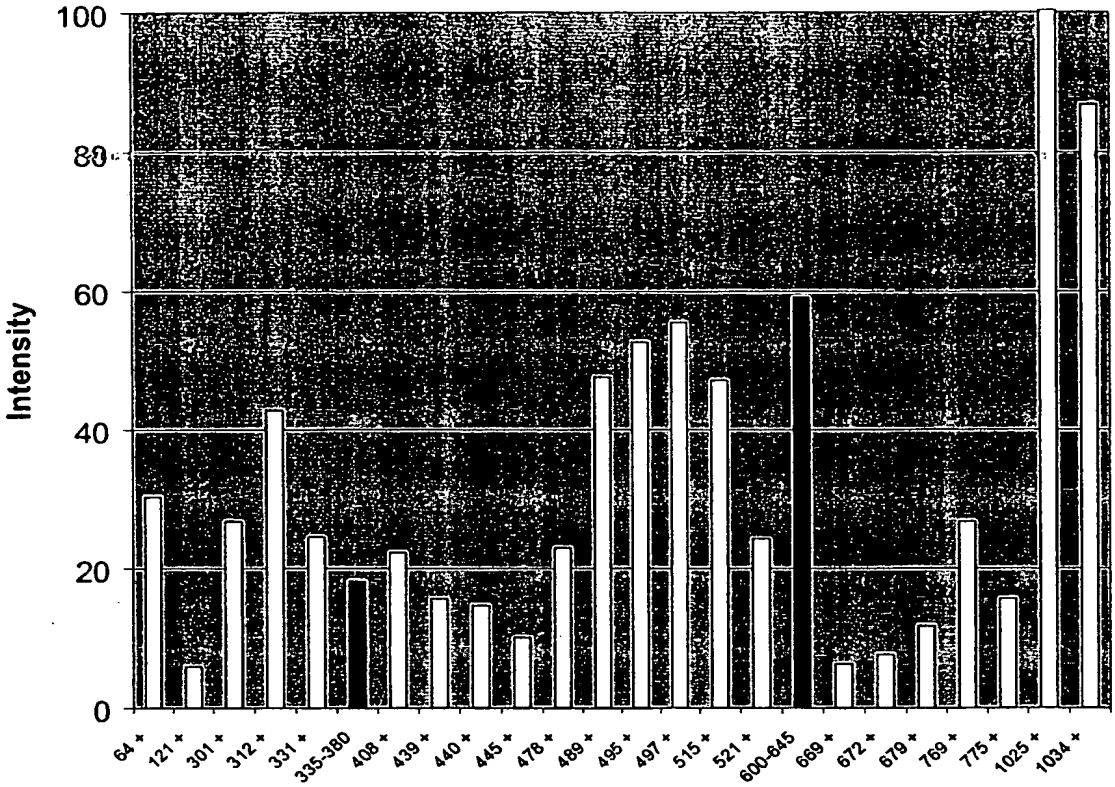
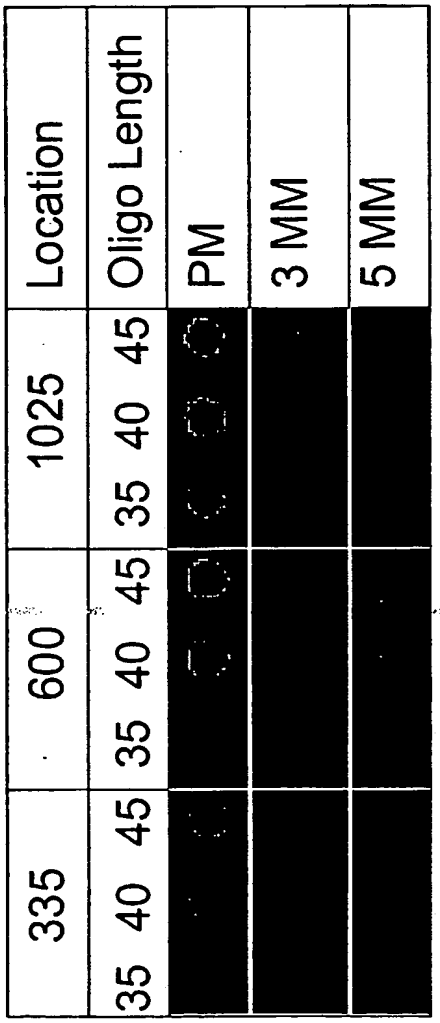


Figure 6



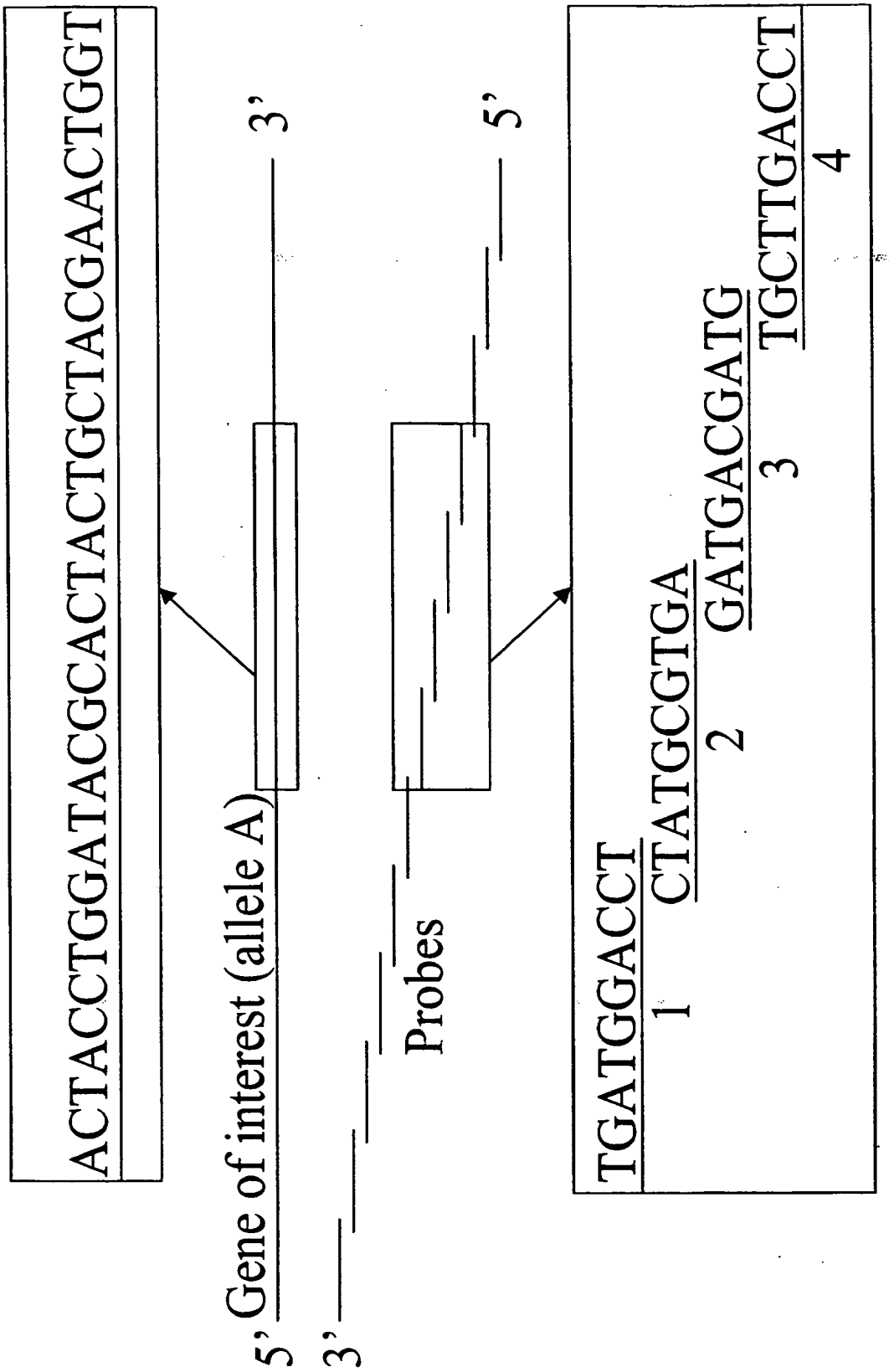


Figure 7

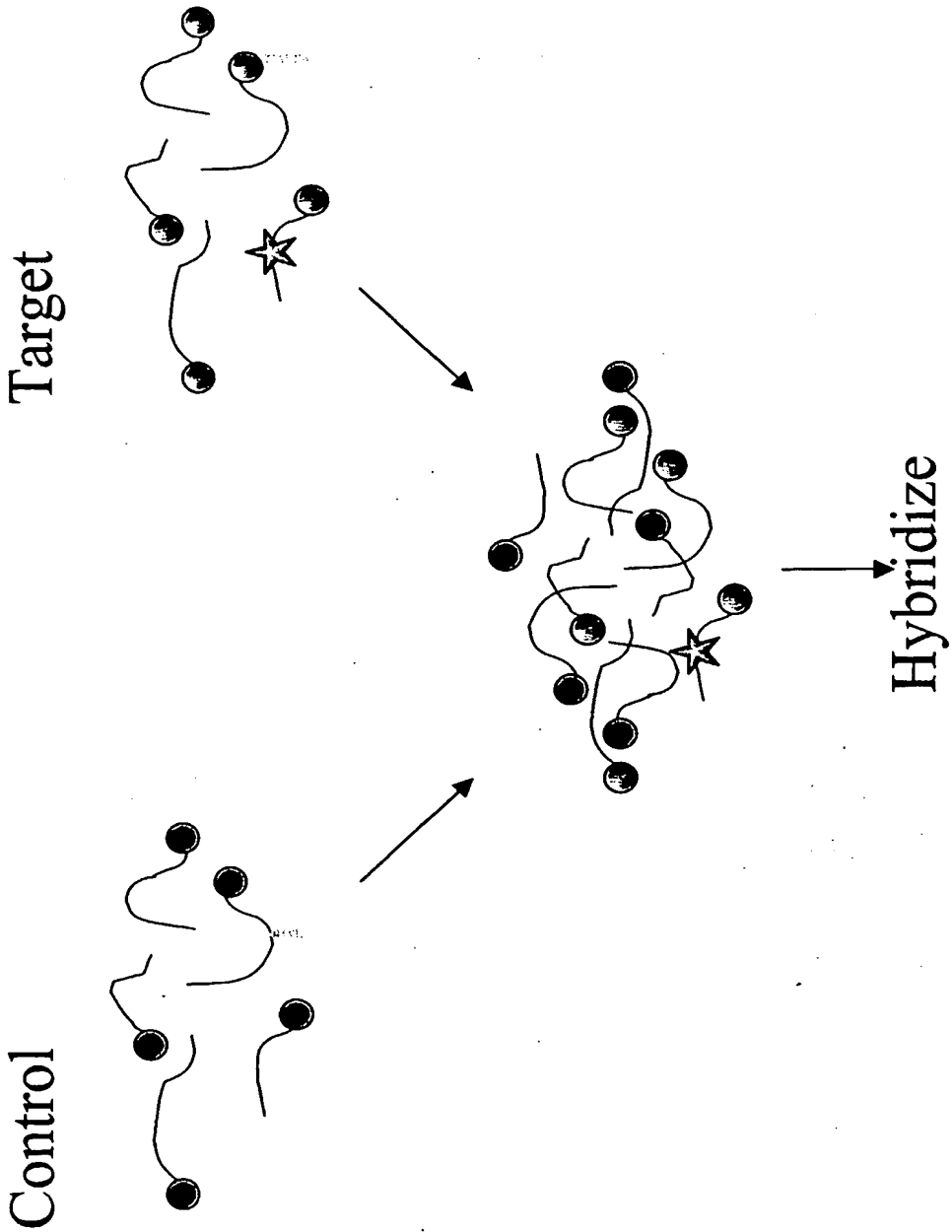
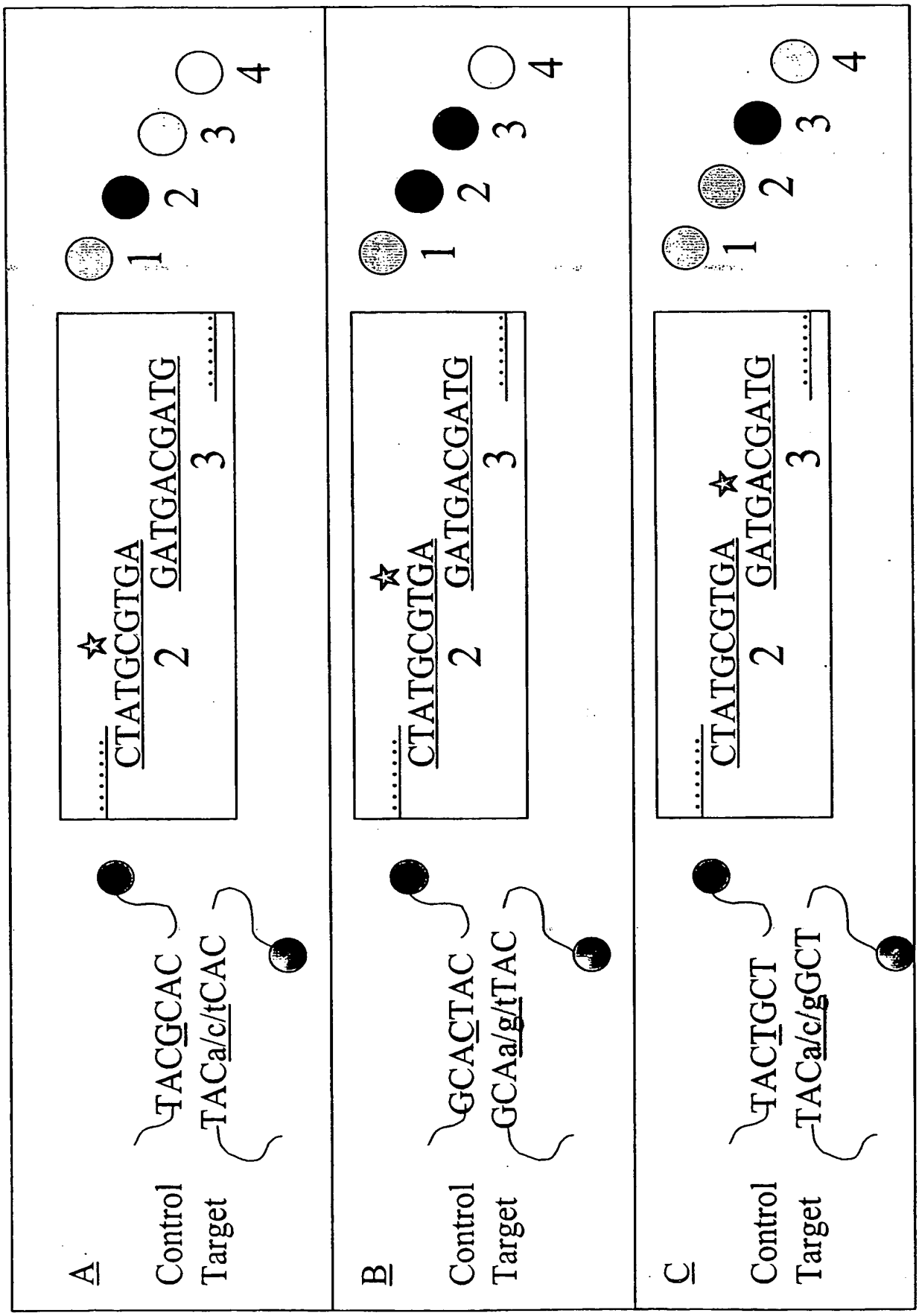


Figure 8

Figure 9



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 September 2001 (13.09.2001)

PCT

(10) International Publication Number
WO 01/066804 A3

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/US01/07775

(22) International Filing Date: 9 March 2001 (09.03.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/521,983 9 March 2000 (09.03.2000) US
09/613,517 10 July 2000 (10.07.2000) US

(71) Applicant: **PROTOGENE LABORATORIES, INC.** [US/US]; 303 Constitution Drive, Menlo Park, CA 94025 (US).

(72) Inventors: **CRONIN, Maureen, T.**; 771 Anderson Drive, Los Altos, CA 94024 (US). **FRUEH, Felix**; 511 Lakeview Way, Emerald Hills, CA 94062 (US). **BRENNAN, Thomas, M.**; 1998 Broadway #1505, San Francisco, CA 94109 (US).

(74) Agent: **HALLUIN, Albert, P.**; Howrey Simon Arnold & White LLP, 301 Ravenswood Avenue, Menlo Park, CA 94025 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) Date of publication of the international search report:
30 May 2003

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/066804 A3

(54) Title: METHODS FOR OPTIMIZING HYBRIDIZATION PERFORMANCE OF POLYNUCLEOTIDE PROBES AND LOCALIZING AND DETECTING SEQUENCE VARIATIONS

(57) Abstract: The present invention relates to a method for optimizing hybridization performance of polynucleotide probes on an array. More specifically, the present invention provides a cost-effective method for designing optimal polynucleotide probes and hybridization conditions to allow simultaneous determination of multiple sequence variations or multiple gene expression levels on an array under a single set of conditions. The present invention also relates to a method of localizing and detecting sequence variations. More specifically, the present invention provides a two-color system for sequence variation localization and detection. The present invention is applicable to high-throughput genotyping of known and unknown polymorphisms and mutations.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 01/07775

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, BIOSIS, MEDLINE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 98 41657 A (CHEE MARK ;AFFYMETRIX INC (US)) 24 September 1998 (1998-09-24) page 5, line 10 - line 30 page 10, line 5 -page 13, line 14 claim 1 ---	1-42
X	WO 97 27317 A (CHEE MARK ;LAI CHAOQIANG (US); LEE DANNY (US); AFFYMETRIX INC (US)) 31 July 1997 (1997-07-31) page 51, line 3 -page 57, line 25 --- -/--	1-42

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 August 2002	13. 11. 2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Aguilera, M

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LOCKHART D J ET AL: "EXPRESSION MONITORING BY HYBRIDIZATION TO HIGH-DENSITY OLIGONUCLEOTIDE ARRAYS" BIO/TECHNOLOGY, NATURE PUBLISHING CO. NEW YORK, US, vol. 14, December 1996 (1996-12), pages 1675-1680, XP002901772 ISSN: 0733-222X page 1680, column 1 ---	1-42
X	CRONIN M T ET AL: "Applying rapid DNA microarray optimization capability to SNP screening and genotyping." AMERICAN JOURNAL OF HUMAN GENETICS, vol. 65, no. 4, October 1999 (1999-10), page A224 XP008006512 49th Annual Meeting of the American Society of Human Genetics;San Francisco, California, USA; October 19-23, 1999 ISSN: 0002-9297 * Abstract * ---	1-42
A	SAMBROOK J ET AL: "MOLECULAR CLONING A LABORATORY MANUAL SECOND EDITION VOLS. 1 2 AND 3" 1989 , SAMBROOK, J., E. F. FRITSCH AND T. MANIATIS , COLD SPRING HARBOR LABORATORY PRESS: COLD SPRING HARBOR, NEW YORK, USA. ILLUS. PAPER. ISBN 0-87969-309-6. 1989 XP002209783 page 11.2 -page 11.19 page 11.45 -page 11.61 ---	1-12, 21-33
A	BLANCHARD A P ET AL: "HIGH-DENSITY OLIGONUCLEOTIDE ARRAYS" BIOSENSORS & BIOELECTRONICS, ELSEVIER SCIENCE PUBLISHERS, BARKING, GB, vol. 11, no. 6/7, 26 April 1996 (1996-04-26), pages 687-690, XP002052193 ISSN: 0956-5663 * whole document * ---	13-20, 34-42
P,X	WO 01 05935 A (ROSETTA INPHARMATICS INC) 25 January 2001 (2001-01-25) page 20, line 20 - line 30; figure 1 page 40, line 22 - line 35 page 69, line 15 -page 70, line 2 --- -/--	1-42

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 01/07775

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	ZHOU YI-XIONG ET AL: "Information processing issues and solutions associated with microarray technology." MICROARRAY BIOCHIP TECHNOLOGY., 17 March 2000 (2000-03-17), pages 167-200, XP008006532 Eaton Publishing 154 E. Central Street, Natick, MA, 01760, USA ISBN: 1-881299-37-6 page 168 -page 169; figure 1 ---	1-42
P,X	DATABASE BIOSIS [Online] BIOSCIENCES INFORMATION SERVICE, PHILADELPHIA, PA, US; 15 September 2000 (2000-09-15) WIEST DEBRA ET AL: "Assessing hybridization specificity in oligonucleotide microarrays." Database accession no. PREV200100494493 XP002209013 abstract & INTERNATIONAL GENOME SEQUENCING AND ANALYSIS CONFERENCE, vol. 12, 2000, page 34 12th International Genome Sequencing and Analysis Conference;Miami Beach, Florida, USA; September 12-15, 2000 ---	1-42
E	US 6 251 588 B1 (DELENSTARR GLENDA C ET AL) 26 June 2001 (2001-06-26) column 2, line 58 -column 4, line 14 -----	1-42

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 01/07775

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sh et)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-42

Remark n Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. Claims: 1-42

Methods for simultaneously determining the presence or absence or two or more sequence variations in target nucleic acids, or for simultaneously monitoring the expression of two or more genes in target nucleic acids, consisting of an iterative process of microarray fabrication, hybridization, analysis and probe optimization.

2. Claims: 43-48

Arrays of polynucleotide probes where the melting temperatures of the probes differ by less than 10 C from the average melting temperature. Arrays of polynucleotide probes where the melting temperatures of the probes differ by less than 10 C. Arrays of polynucleotide probes where the melting temperatures of the probes differ by less than 5 C.

3. Claims: 49-63

Methods for simultaneously determining the presence or absence of a sequence variation in a target nucleic acid consisting in the differential labelling of a control and a target nucleic acid sequence, simultaneous hybridization to a polynucleotide array and analysis of pattern differences.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 01/07775

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
WO 9841657	A	24-09-1998	EP	0972078 A1	19-01-2000
			WO	9841657 A1	24-09-1998
			WO	9939004 A1	05-08-1999
WO 9727317	A	31-07-1997	AU	2253397 A	20-08-1997
			EP	0880598 A1	02-12-1998
			JP	2002515738 T	28-05-2002
			WO	9727317 A1	31-07-1997
			US	6344316 B1	05-02-2002
WO 0105935	A	25-01-2001	AU	6213500 A	05-02-2001
			AU	6213600 A	05-02-2001
			EP	1200820 A2	02-05-2002
			EP	1200625 A1	02-05-2002
			WO	0105935 A2	25-01-2001
			WO	0106013 A1	25-01-2001
US 6251588	B1	26-06-2001	NONE		